UNIVERSITAT DE BARCELONA

Characterization of Designated Communities of Geospatial Legacy Information and Their Application in Appraisal and Digital Preservation Decisions: A Case Study

Anita E. Locher

UNIVERSITAT DE BARCELONA

Department of Library and Information Science and Audiovisual Communication

Doctorat en Informació i Documentació en la Societat del Coneixement

Characterization of Designated Communities of Geospatial Legacy Information and Their Application in Appraisal and Digital Preservation Decisions: A Case Study

A thesis submitted by **Anita E. Locher** in partial fulfilment of the requirements of the University of Barcelona for the degree of

Doctora per la Universitat de Barcelona

Director: **Dr. Miquel Términs Graells** (University of Barcelona)
Co-director: **Dr. Ross Purves** (University of Zürich)

Barcelona, May 2017

to my daughter Yuna

Acknowledgements

I want to thank all the souls who motivated and accompanied me on this journey, my friends and colleagues who spent their time and heart on this project and all the participants and interview partners who so kindly offered their time and expertise. My special thanks go to Grelda, Eli and Dani, without whom this thesis would not have been possible. Finally, I want to thank my supervisors for their guidance.

Abstract

Institutions that recognize the value of digital geographic legacy data for future use develop the motivation to preserve the data for the long term. International standards on preservation, such as for an open archival information system (OAIS), require having one or more designated communities identified for whom to preserve the data to adjust preservation measures to the communities' needs. Lacking scientific literature about designated communities and the way to identify them, we use qualitative research methods to explore user groups and their needs. Our goals are to build basic knowledge on western culture user communities of legacy geodata (goal one) and to use this knowledge to adapt existing preservation practice, specifically appraisal, of the Institute for Cartography and Geology in Catalonia (ICGC), a geodata producer and service provider (goal two). The Delphi method is used to reach goal one and through focus group interviews we get to know current legacy geodata users in Catalonia, which allows us to draw conclusions for goal two.

As results, we get characteristics of thirteen user types of different professional profiles. We show that they can be clustered by the similarities of their needs into six groups, of which four are relevant to a long-term archive: geographers, historians, urban planners and the general public. Through interviewing members of these four clusters, we extracted characteristics that are mentioned by various preservation guidelines as relevant for appraisal and selection of geodata. The characteristics are the interaction type with the data (interaction type), the data age range users are interested in (time range), their technological knowledge and in particular knowledge of geographic information systems (GIS knowledge), the maximal time span between data acquisition they could work with (acquisition interval), the preferred type of product, scales and file formats and the significance of the user group. The resulting profiles of the Catalonian users were employed to recommend a relevant designated community to the ICGC and to illustrate how that might influence detailed appraisal decisions for part of their production. Users input also allowed us to evaluate the importance of existing appraisal criteria for geodata.

Thanks to triangulation of the Delphi study results with the user input we can conclude that the Delphi method is an efficient way to predict legacy geodata user profile characteristics. Nevertheless, direct user interaction reveals far more details that can influence preservation decisions. Practical application of the profiles showed that some modules were not useful

for appraisal; e.g., file formats. By applying what we have learned about user needs to appraisal criteria and other preservation decisions we give the user a voice that will compete in the final decisions about implementations of user-adapted legacy geodata preservation with technical, organizational and legal aspects.

Resumen

La motivación de las instituciones que reconocen el valor de datos geográficos heredados para su uso futuro es poder preservarlos a largo plazo. Estándares internacionales de preservación como el 'Open archival information system (OAIS)' requieren que se identifiquen uno o varios grupos de usuarios potenciales (comunidades designadas) para los que se van a preservar los datos y a cuyas necesidades se ajusten las medidas de preservación. En ausencia de literatura científica sobre las comunidades designadas y la manera de definirlas, usamos métodos de investigación cualitativos para explorar grupos de usuarios y sus necesidades. Las metas de este trabajo son la creación de conocimientos básicos sobre comunidades de usuarios de datos geográficos heredados de la cultura occidental (meta número uno). El uso de este conocimiento para adaptar prácticas de preservación existentes del Institut Cartogràfic i Geològic de Catalunya (ICGC), un productor y proveedor de servicios de geodatos, en particular la valoración de archivos, es la meta número dos. El método Delphi se usó para alcanzar la meta número uno y a través de entrevistas en grupos focales conocimos usuarios de datos geográficos heredados en Cataluña, lo que permitió sacar conclusiones para la meta número dos.

Como resultado, obtenemos las características de trece tipos de usuarios con distintos perfiles profesionales. Demostramos que se pueden agrupar en seis grupos gracias a la similitud de sus necesidades. Cuatro de estos grupos son relevantes para un archivo a largo plazo: geógrafos, historiadores, planificadores urbanos y el público en general. Entrevistando miembros de estos cuatro grupos obtuvimos características que son mencionadas por varias directrices de preservación digital como relevantes para la valoración y la selección de datos geográficos. Estas características son: el tipo de interacción con los datos (tipo de interacción), el rango temporal de los datos que interesa a los usuarios (rango temporal), su nivel de conocimiento tecnológico, en particular su conocimiento de sistemas de información geográfica (conocimiento SIG), el intervalo de tiempo máximo entre dos adquisiciones de datos por parte del archivo con el que puede trabajar el usuario (intervalo de adquisición), el tipo de producto, la escala y los formatos de fichero preferidos y, finalmente, la importancia del grupo. Los perfiles resultantes de los usuarios catalanes se usaron para recomendar una comunidad designada al ICGC y para mostrar, en parte de la producción del ICGC, cómo la comunidad puede influir en decisiones de valoración concretas. Las respuestas de los usuarios también permitieron la evaluación de la importancia de los criterios de valoración para datos geográficos existentes.

Gracias a la triangulación de datos entre los resultados del estudio Delphi y las respuestas de las entrevistas pudimos concluir que el método Delphi es eficiente para predecir características de perfiles de usuarios de datos geográficos heredados. Sin embargo, la interacción directa con el usuario reveló más detalles que pueden influir decisiones relacionadas a la preservación digital. La aplicación práctica de los perfiles mostró que algunos módulos no fueron útiles para la valoración de archivos; por ejemplo los formatos de ficheros. Aplicando lo que aprendimos sobre las necesidades de los usuarios a los criterios de valoración de archivos y otras decisiones de preservación dimos voz a los usuarios. Esta voz competirá en la decisión final sobre la implementación de medidas de preservación guiadas por el usuario con argumentos técnicos, organizativos y legales.

1 Introduction

1.1 Importance of the topic

Preservation of geographic data is important because geodata can be combined with information on various topics, such as biology, sociology, the economy and many others, to create new understanding of other subjects. The European Commission assumes complete new possibilities with digital data, such as being able to combine data sets of different disciplines, thanks to their possible interoperability, easier and faster analysis and international availability hardly achievable with analogue information (European Commission. High Level Expert Group on Scientific Data, 2010). These predictions made in the *Riding the Wave* report are said to be valid for all sciences including the geosciences. As long as legacy geodata are digital, they have the same combining potential and are granted the same value as recent data. Nevertheless, the effective combination of data sets or the smooth analysis of time series is only possible with homogeneous semantics over time and between producers. Harmonisation efforts in the geospatial sector are now going on through the Infrastructure for Spatial Information in the European Community (INSPIRE) initiative.[1]

Since geographic information is produced in digital form, it is also stored this way. Cartographic production is a professional field in the area of geography, affected by the digital shift and now concerned about the long-term survival of its data. Digital preservation is dedicated to maintaining digital information with long-term value, keeping it readable for future access while guaranteeing its authenticity and reliability. Preservation tasks have classically been handled by libraries, archives and museums. Well-established preservation systems and processes developed for analogue media have had to be transformed to adapt to the new digital context because, for example, digital systems allow remnants of production to accumulate invisibly and copies and versions to reach almost unmanageable complexity. Government bodies and other institutions are now faced with a huge amount of

[1] Website of INSPIRE: http://inspire.ec.europa.eu/

data and with questions about what should be kept and what can be disposed of. New interpretations of law about archiving are necessary because digital production has blurred the frontier between published and non-published documents that until now has been decisive for the attribution of curation responsibility.[2] In this grey zone of responsibility over legacy data, new stakeholders can step in. Geodata producers that recognize the value of their legacy holdings can start to preserve them with or without legal obligation to do so.

Libraries and archives have always practiced selection and appraisal of documents susceptible to be part of their collections. On the one hand, not everything gains historic value over time or helps to understand the functioning of the producing institution. And on the other hand, it is simply not sustainable for an institution to preserve everything in the long term, not even for an archive. The sheer amount of data, the complexity of dynamic geospatial databases and the high density of information, especially in raster files, turns preservation into a big challenge. To select relevant documents for long-term preservation, libraries and archives use different systems: libraries are guided by their collection development policies, which are intended to serve current users, while archives preserve for unknown future users. Public archives need a legal incentive to preserve and select documents with historic value for potential future use. With the advent of data warehouses and digital repositories, the research value of legacy geodata has come to the fore. Commercially oriented repositories, especially, will search for tangible research value reflected in data reuse and are therefore interested in knowing about the needs of the user community now and in the future. Palmer et al. express it this way:

> *[...] with digital data, assessments of value will not only require understanding of the content but also its structural and semantic make-up in relation to how analysis will be performed by various service communities* (Palmer, Weber, & Cragin, 2011).

Because the reason to keep a document is always potential reuse, user opinions are involved in selection in various ways. In the case of the archive, stakeholders interested in preserving digital assets have developed standards that involve the user under the concept of the 'designated community'. This concept, as we will see further on, plays a significant role at various points in managing long-term archives (see chapter 2.2.3). The designated

[2] http://www.unesco.org/webworld/memory/legaldep.htm#Formoflegislation

community concept appeared first in one of the most respected standards in digital preservation: the reference model for an open archival information system, abbreviated OAIS. 'OAIS' stands for the standard but also for an archival system that was implemented in accordance with the standard. For clarification, we use 'OAIS' when the archival system is meant and 'OAIS reference model' when talking about the standard. The OAIS reference model defines an archive as 'an institution that intends to preserve information for access and use by a Designated Community'. A second standard in digital preservation that addresses the designated community is an audit guideline that checks conformity of a digital repository with the OAIS reference model. It is ISO 16363, the Trusted Digital Repository Checklist. Obviously, just as the designated community is essential for correct long-term archiving according to the OAIS reference model, it is as well for the auditing guideline.

The OAIS reference model, which requires the designated community be defined, was created to identify an ideal system. At the same time, it was kept as general as possible to fit different implementations, though few experiences of implementations are documented. Indeed, in a speech at the Future Perfect conference, Rothenberg talked about the few real OAIS implementations and how institutions struggle with individual aspects of the model (Rothenberg, 2012). Institutions that want to implement OAIS face many challenges that come up when confronted with the constraints of reality. Institutions with fewer resources that want to preserve digital holdings need practical, ready-to-use guidelines and might implement only part of the standard. Nonetheless, the definition of the designated community is so fundamental to the European Long Term Data Preservation Common Guidelines that it is recommended as one of the first steps in implementing a preservation workflow (Ground Segment Coordination Body, 2012).

While libraries historically have wanted to constantly increase access, archives were much more restricted by the contradiction of conservation measures and the potentially damaging use of unique documents. Not so in the digital world, where access does not damage the file and the original can be duplicated easily. Digital archives can now bring the user to the fore. Even though more intense user involvement might exist in practice, there is a lack of scientific literature about archive users in general and about the designated community in particular.

Responding to rising interest in user-centred research and the lack of scientific literature, this thesis wants to increase understanding of the nature of designated communities. The documentation and analysis of the research process should be a step in the direction of finding a method to define designated communities of other data. Taking into account the value of legacy geodata, we use the knowledge about legacy geodata users to face an existing problem about geographic data preservation. The topic of this thesis is represented in the following figure as the intersection of three fields of science with their proper interest.

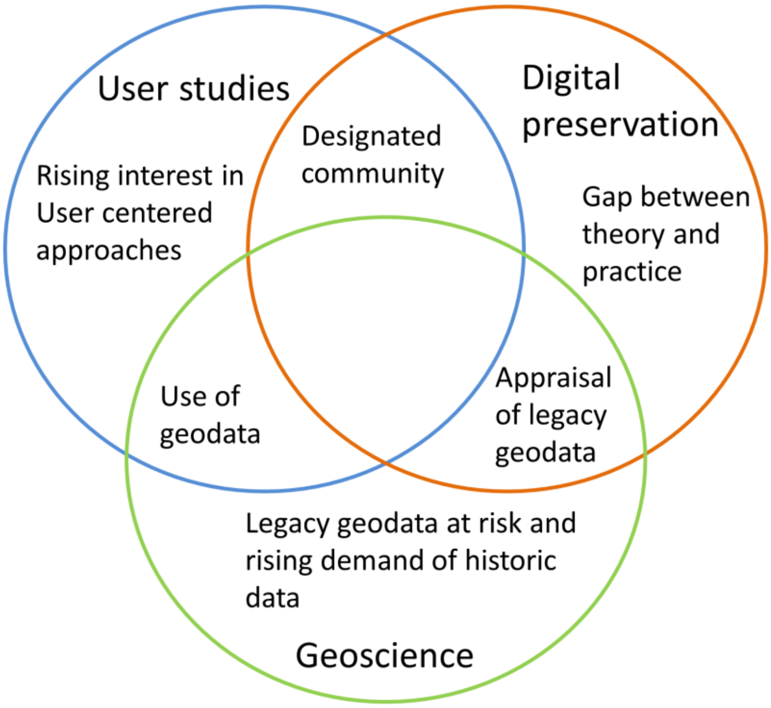
User studies
Rising interest in User centered approaches
Designated community
Digital preservation
Gap between theory and practice
Use of geodata
Appraisal of legacy geodata
Legacy geodata at risk and rising demand of historic data
Geoscience

Figure 1: The three fields that influence this research are user studies, digital preservation and geosciences.

Each circle represents a field of study that is not directly connected with the others: user studies, digital preservation and geosciences. Each field adds a characteristic to the importance of this research: geodata are at risk of being lost, overwritten or poorly managed; user-centred approaches and user-experience testing are becoming increasingly popular (Dent Goodman, 2011) and there is a lack of guidance and research around the concept of the designated community in digital preservation that this research intends to approach. At the intersection between user studies and geosciences is the study of geodata use. Where geosciences and digital preservation meet, we find the appraisal and preservation of geospatial data. At the intersection of digital preservation and user studies lie the definition of the designated communities and the research of their needs. This thesis is located at the intersection of all three circles. As a first step, we want to define the user profiles for users of legacy geographic information so that geodata archives can choose their designated communities based on the developed knowledge about those users. As a second step, we want to give users a voice in the appraisal of existing geodata.

1.1.1 Motivation

The idea for this study came about when talking to members of a local geodata producer and its library staff about the fact that they could not define who uses their historic content. Further interviews showed that they are not alone with this concern and others would like to know more about their users as well. This interest is not randomised, because knowing the user can help decision-making in digital preservation, especially for defining the designated community. Identification is especially difficult because decisions tailored to the current designated community can negatively affect future user groups. Despite the importance of the designated community, archives lack guidelines on how to identify it; to our knowledge there are only two scientific texts explaining a way of defining the designated community. One is a theoretical instruction to define the designated community’s knowledge base by modules (Giaretta, 2011) and the other is an implementation that demarcates user groups based on the frequency of use and other properties measured with web analytics (Kärberg, 2014). We will write about them in chapter 1.2.2.

A second incentive for this study was the need of the geodata producer that served as the study case to know what geospatial data to keep in its holdings and what to dispose of. Selection and appraisal are nothing new to archives (Craig, 2004), though digital data and in particular the complexity of geographic information units and their interdependence increases the challenge of appraisal. Most geodata preservation projects have approached this challenge by revising and discussing existing appraisal criteria for geodata or by creating new recommendations, which we are going to analyse in chapter 3.2.1.

1.1.2 Why a user-centred approach?

In 2002, the U.S. National Research Council published appraisal criteria for the geological sector inspired by the guidance of the National Archives and Records Administration (NARA). One of the criteria states that worthiness of long-term preservation rises if data have potential applications. The National Research Council (Committee on the Preservation of Geoscience Data and Collections, Committee on Earth Resources, National Research Council, & National Academy of Sciences. National Research Council, 2002) wanted this criterion to be evaluated by an evaluation board containing the data users. Later recommendations by the U.S. National Oceanic and Atmospheric Administration's (NOAA) and the Library of Congress also called for user input to the appraisal process (S. P. Morris, 2010; The National Oceanic and Atmospheric Administration [NOAA], 2007, 2008). In almost the same manner, Abrams et al. plead for considering use cases when appraising geodata. He says in a meeting about a national strategy for appraisal and selection of geodata in the United States:

> *Appraisal is still largely done on a domain-specific basis without adequately taking into consideration appraisals for re-use purposes that cross domains. Consider ways to include perspectives on future use and analysis of geospatial data.* (Abrams et al., 2010)

Abrams' call leads us to consider not only classical appraisal criteria such as historic value, but immediate scientific value to other disciplines. Despite all these calls for user involvement, the available literature does not explain how or to what point users have influenced the definition of existing criteria. Due to the central role of the designated community in the two mentioned preservation standards, we think user involvement deserves a place in appraisal. This study wants to hear from users in order to aid improvement of geodata appraisal and will document the user involvement for the sake of transferability of the method.

1.1.3 Why geospatial data?

It must be said that a user-centred approach could be applied to the appraisal of many different kinds of digital data. We chose geodata for the following reasons:

- Geodata is currently a focus of business because of its increasing popularity. In 2001, the GeoConnections network had already detected a growing interest in georeferenced data and increasing awareness of its value (Sears, 2001). Now, years later, we can see the repercussions in the market, which offers many applications using georeferenced data. A report to the European Commission on the reuse of public sector information by companies (Vickery, 2011) shows that 30% of the activities with public sector information use geographic data. With new devices that can read, create and display it, creation of spatial data also increases. Online applications such as OpenStreet maps (Haklay & Weber, 2008), Google My Maps and similar map-sharing platforms lead to participative and spontaneous map creation (Kanehira, Arakawa, Yasumoto, & Wada, 2016) and location-based services combined with the location awareness of mobile phones call the attention of industry players (Mountain & MacFarlane, 2007).

- Geospatial data are at risk. As more devices use and capture spatial data the technological forms disperse, the production is decentralised and more producers should be sensitised for preservation actions (Mcgarva, Morris, & Janée, 2009). The sheer amount of production exceeds the storage capacity of individual institutions (Sweetkind, Larsgaard, & Erwin, 2006).
- Spatial information has the potential to be combined with data from many other sectors, which increases its value. Indeed, many data sets only develop full comprehensiveness when spatial reference data underlay them (Bos, Gollin, Gerber, Leuthold, & Meyer, 2010; Lutz & Kolas, 2007). Demand by other fields of science has gone so far for the U.S. National Aeronautics and Space Administration (NASA) to maintain a Socioeconomic Data and Application Center.[3]

1.2 Definitions

We shall introduce some concepts used in this thesis to avoid misunderstanding and to clarify terms that are used by various scientific communities or domains. Because we refer to the digital preservation standards ISO 14721 and ISO 16363 in this thesis and use them in practice we will also employ their terminology. Because these standards do not cover the thematic terms specific to geosciences, additional definitions are given.

[3] http://sedac.ciesin.columbia.edu/

1.2.1 Definition and differences between the terms user, consumer and designated community

Let us shed some light on the terms used in the digital preservation standards ISO 14721 and ISO 16363 that have to do with the user: ‘user’, ‘consumer’, ‘customer’ and ‘designated community’. In library and information science, we talk about the user when referring to a person who makes use of services at the library or archive. When the service is transferable or when speaking about a product, the consumer is the buyer and might not be the end user. Economically speaking, we more often hear ‘the consumer’ in phrases such as: ‘the consumer is at the source of demand.’ Market studies want to target the consumer, who is the one who makes the financial decisions and interacts with the service, and not necessarily the end user. In a group of people where one buys for a community, all members are considered consumers. All the potential consumers together form the market. Even though in information science the term ‘consumer’ is not common, we will use it when appropriate with this slight difference in meaning from ‘user’. The consumer likes to see himself or herself as the customer. For maintaining a good relationship with the consumer, companies and institutions use the word customer when in direct contact with the consumer. In keeping with this practice, the OAIS model uses labels such as ‘customer comments’ and ‘customer services’ instead of consumer comments or services.

The designated community is an uncommon concept that is different from the user or consumer; it appears only in the context of digital archives. It was first used in ISO standard 14721 and was inherited by ISO standard 16363. The designated community is defined as

> *An identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the archive and this definition may change over time* (Consultative Committee for Space Data Systems [CCSDS], 2012).

Therefore, not all consumers of archived data must belong to the designated community; there might be members of the designated community who will not use the data. On the one hand, an archive wants to define the designated community in a way to get as many consumers as they can and to deny services to as few as possible. On the other hand, it must maintain economic sustainability and therefore limit its services.

While the terms user and consumer imply those who currently make use of services at the archive, the designated community implies a vision into the future. The long-term mission of the archive is transmitted to a long-term commitment to the designated community. The concept of designated community therefore reaches from the current community to all members of this community in the future. Current users as well as future users of this community must be able to read and understand the information. This does not mean that the definition of the designated community cannot change, but actions tailored to the current needs and knowledge base of the designated community should not impede satisfaction of the needs of future members of the same community.

The concept of 'client' is used in the OAIS model only with the technological meaning of client server or client system and will be used here in the same way.

1.2.2 The user in information science

Little scientific literature is available about the concept of designated community. Therefore, we will take a broader approach by shedding light on user studies in libraries and archives in general, as well as addressing the role of a designated community.

Libraries are committed to the users who they intend to serve. The International Federation of Library Associations and Institutions (IFLA) recommends taking users into account when developing a collection development policy:

> *Ideally, the compilation of the document requires the active participation of both users and administrators, thereby improving communication between the library and its clientele. The policy statement serves as a contract with the library's users; it has the function to demonstrate to individuals within an institution what they can expect of the library both in form of collections and of services.* (Biblarz, Tarin, Vickery, & Bakker, 2001).

Libraries and archives have shown and continue to show interest in their users to develop or revise their services. Studies from the 1940s to the 1960s concentrated on such things as better shelving and loan statistics, or intended to identify hours with high user afflux (Wilson, 2008). Still today, many internal user studies without scientific aspirations repeat this scheme. At the Royal Society's 1948 Scientific Information Conference there was one exception though; J.D. Bernal presented on 'what [working scientists] read, why they read it and what use they made of the information' (Bernal, 1948). The new chapter on information needs in the journal *Annual Review of Information Science and Technology* in the mid-1960s became a platform for publishing about user needs. In the 1970s the first academic research on users was positivist and focused on subscription rates, occupation and social class (Wilson, 2008). More recently, qualitative methods have become more popular and even dominate the field. Currently research is interested in which information resources the readers use and how they access them. This knowledge helps libraries to adapt their acquisition policies and distribution of financial resources. Dent Goodman defends user-centred approaches: 'ideally, library-related research should somehow be connected to user need' (Dent Goodman, 2011). As user needs evolve, we need to monitor information seeking and use constantly to adapt services to this changing environment.

1.2.2.1 The user in the selection process

The user also has a role in the selection and appraisal process of libraries and archives. Libraries have various approaches in acquisition that reach from least to maximal user involvement. The user is less involved when expert acquisition is practiced, where only library staff decides what to acquire for the collection, and the user might only be able to make acquisition suggestions. An intermediate user involvement would be assisted acquisition by topic experts. These are subject experts who suggest a certain number of titles in their area suitable for the library. Experts work for the library but also represent specific user segments. The biggest role is given to the user in patron-driven acquisition, also called demand-driven acquisition. The purpose of this new book selection model is to avoid spending resources on books that will never be read, a response to increasing pressure on library budgets and rising interlibrary loan numbers. It consists of a system that permits library users to order books of their choice for the library holdings. The model was tested in academic libraries where the university staff selected the books to buy from a vendor's catalogue. It was established in 1999 in some universities of the United States of America but due to weaknesses in the business plan was abandoned in 2006; indeed, libraries encountered an increase in acquisition costs. New models allow libraries to better control the budget. Today librarians might preselect a universe of titles from which users can choose. Patrons have a limited budget to spend on those items. Patron-driven acquisition works particularly well for ebooks when the availability to the user is almost immediate, so that we can say that patron-driven acquisition in its recent form is relatively new.

There is no similar system for selecting archival material for long-term preservation. For obvious reasons, patron-driven acquisition cannot be applied to archives, but the development in acquisition processes in libraries shows the interest that is given to how the material is used. This interest is present also in archival services and expressed in the OAIS reference model and in ISO standard 16363 through the concept of the designated community.

1.2.2.2 The role of the designated community

Archives are challenged to save their collections for an uncertain future user. Their way of involving the user in collection development is by defining their designated communities. It is no secret that digital preservation is a costly endeavour. If it were not for limited resources, an archive would save all data and provide services to all users with all levels of knowledge about the data. Choosing one or more user groups as a designated community is part of a series of measures that should guarantee the sustainability of long-term preservation throughout fluctuation of resources. Measures such as the limitation of technologies handled by an archive reduce description, migration and eventual rendering costs. Smart election of technologies reduces the frequency with which preservation measures such as format migration must be applied. In the same way, an archive can limit its services to a designated community instead of trying to serve all consumers. Moreover, providing a clear idea of who makes up an archive’s designated community can improve understanding of the archival material and promote shared responsibility over collection development between archives. Collaboration also adds to the economic sustainability of preservation measures (Walters & Skinner, 2010). Networked archives can direct collections offered to them to other archival institutions they know are interested in these collections or point users that are not part of their designated community to institutions that do provide services for them. Archives with similar interests should therefore come to an agreement on the designated communities they serve and the collections they hold, in order to reduce their services to a sustainable size.

Information stored in the OAIS should be accessible to, understandable by and useable by the designated community. The designated community has a certain knowledge base, which allows it to translate the data into more meaningful information. The OAIS reference model requires the archive to preserve all information needed by the designated community together with the data. The information needed to represent and understand the data depends on the designated community's knowledge base. Therefore, the OAIS must understand the knowledge base of its designated community to understand the minimum information that must be maintained. The community itself, its service requirements and its knowledge base can change with time. The community can, for example, lose its familiarity with certain terminology.

This is why the designated community has to be monitored and metadata that should guarantee understandability and usability has to be periodically revised and eventually expanded to add additional explanation. Understandability is the capacity of data and metadata to 'explain themselves' to a human user – its provenance and purpose, technical specifications and intellectual content. Usability is a wider concept that encompasses not only intellectual understanding but also technical and administrative feasibility such as the needed effort to render data, integrate them into a modern system and compare them to other data. The International Organization for Standardization defines usability as '[t]he extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use'. An archive has to know who its designated community is to achieve understandability, but to achieve usability it must also know the community's goals and the tasks it performs on the data. The OAIS should ensure that the designated community can understand the data without having to call in experts from the producer side. The OAIS reference model proposes that archives monitor the designated community by surveys, via a periodic formal review process, via community workshops where feedback is solicited or by individual interactions (Consultative Committee for Space Data Systems [CCSDS], 2012).

1.2.3 The user of geographic information

We are all users of geographic information. We search for a restaurant nearby, locate the public transportation that brings us to point B, search for the location of a meeting point, shop or a friend's house, orient ourselves in a town with a city map and find a route to another town with a navigation system. All these tasks are performed increasingly in digital ways, because the number of digital resources is growing; most people have mobile computers (smartphones included) and broadband networks are widely available. This new context has allowed new applications, for example for location-based services. Location is important not only for the individual but for science and administration. Government and the private sector have to take location into account when acting on policies, strategies and plans (Longley, Goodchild, Maguire, & Rhind, 2015). The fact that map-based information assists orientation, speeds up navigation and can reveal facts that are hard to discover in the mass of textual data makes geodata in the form of a map or plan so attractive. The easy collection of geodata in a handheld device, such as GPS, and the widespread tools that can read that data make geographic concepts more and more commonplace. While online services and web-based mapping tools reach a wider public, many map users do not only consult but also amend maps or create maps themselves. Location-based services and our own geodata collections mostly still depend on basic geographic information as background reference.

The previous examples describe uses of recent geodata. Exploration of historic geodata is much more rare and we might not even remember doing it – the day we looked at the our childhood pictures and tried to locate where they were taken, the exhibition about climate change that showed the glaciers from then and now, the book about our hometown with the drawing of the city centre as it was long ago – these are all examples of casual encounters with historic geodata that we might be familiar with. In addition, there are professional uses and leisure activities that require consultation of legacy geodata that we might not think of and are going to explore in this thesis.

1.2.3.1 User studies in geoscience

We performed a literature review on map use to inform this study about the purpose and intensity of geodata use. In the library world, the LIBER group of map libraries conducted a survey of its users (Millea, 2005). Remarkable about this study is that it reused a questionnaire that had been employed in 1987-1988, which enabled it to reflect changes in map use. Informative for our study are the categories of use and the geographic areas required. The most prominent category in both editions of the survey was 'history', with 30% of the users signalling this subject of study. Cartography made it to the next most popular category. Millea explains that many of the respondents who used the category 'other' stated that their subject was related to the environment. Another emergent field was 'urban geography'. While both topics were not available in the previous version of the survey and therefore were not mentioned in the report, other categories from the first survey were left blank in the second edition. This shows changing and emerging research trends. Regarding the geographic areas required, we come to know only that oversea map requirements were prominent and that this was already the case in the first survey in 1987-1988. Additionally, both surveys report high rates of reproduction requests (by photocopy, photography and scan), users requiring assistance in their inquiries (80%) and high use of reference material. While these reports give insight in the user categories they are not sufficiently detailed to answer questions about which type of product the users would prefer and why they would use certain geographic products and not others. More detailed information on the purpose of use is desired to accomplish the goals of this thesis (see chapter 1.3).

A similar literature review in the geosciences revealed papers concerned with the improvement of map design (Stigmar & Harrie, 2011; Suchan & Brewer, 2000). This trend finds continuation in usability testing for interactive digital maps (Harrie, Stigmar, & Djordjevic, 2015; Kramers, 2008). Another branch applies user-centred approaches but is not necessarily interested in the interaction between user and map; for example, Bielecka et al. investigate data quality by means of a user study (Bielecka, Leszczyńska, & J Halls, 2014), while Elzakker and van de Berg (Elzakker & van de Berg, 2010) use focus groups and surveys to optimise the generalisation process of topographic maps. Finally, in his dissertation Elzakker describes map use related to the purpose of the map itself (Elzakker, 2004) and introduces MacEachrens map use cube (MacEachren & Kraak, 1997) (see Figure 2).

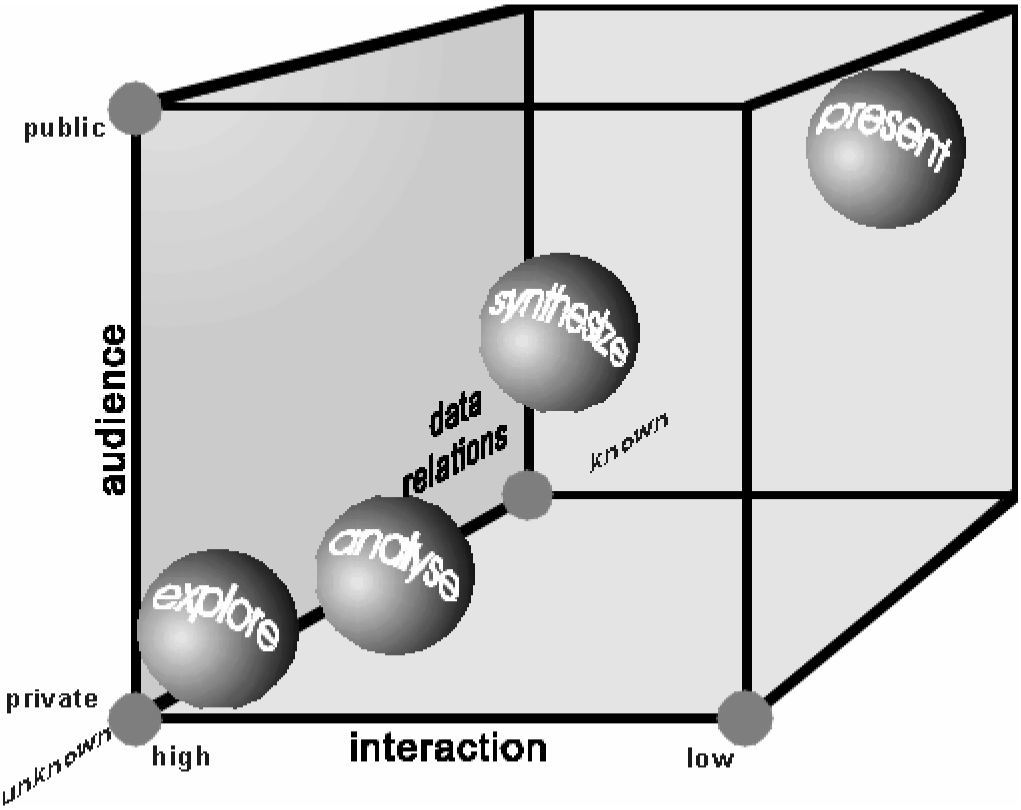

Figure 2: MacEachrens map use cube with the four map use goals placed in the three-dimensional space of the cube: explore, analyse, synthesize and present. Source: (Elzakker, n.d.)

The position on the axis of the cube answers the following questions:

- Is the map private (created by a single user for his or her own use) or public (created to satisfy the highest number of needs possible)? Vertical axis.
- Is there a high human interaction with the map, such as in representations with underlying geodatabases, or is interaction low, such as in rasterized maps where the user cannot change the presentation? Horizontal axis.
- Should the map reveal unknowns or present knowns? Z-axis.

MacEachren & Kraak also define four map use goals that can be situated in the three-dimensional space of the cube. They use the term visualization for maps with high human interaction that reveal unknown patterns to individual users. The goal ‘present’ refers to presenting existing knowledge and is situated in the visualization corner while ‘exploring’ is placed in the opposite corner. Data exploring is understood as a way of querying that reveals unknown patterns in geodata. MacEachren & Kraak’s model of map use goals has been useful to this thesis because it inspired the questions about the ‘interaction type’ (see the user modules in chapter 5.1.1). However, the data use cube is a theoretical model that does not report real data use. And Elzakker's thesis focuses only on explorative data use and ignores simple consultation. Additionally, he does not distinguish between current and legacy data. For these reasons, neither of these scientific texts can answer our research question, and we still lack an understanding of how legacy geodata are used and for what purpose.

Apart from scientific literature, there are diverse other sources of information available to learn something about the user, to include requests to the producer and web statistics. If data are protected by copyright and the producers are the only ones that sell the data, they will know exactly how many requests for current and legacy data are deposited and which bodies require them. When data are available for free on the internet and in analogue form through the library, we have to resort to other sources such as visits to the homepage, downloads, library loans and requests and entrance statistics. The degree of knowledge about who the user is and the purpose of use decreases drastically as compared to paid and eventually even tailored service to a customer. Additionally, where media mediators are at work, such as people reusing data for further publication or introduction in another information system, the end user or consumer of the data is completely unknown to the producer. Usually there is an intermediate level of knowledge about data reuse when consultation is free of charge but permission is required for republication. In cases of formalised data requests, the purpose, data size, data characteristics and extend of reuse are exchanged and could inform the knowledge base of the geodata user.

1.2.4 Definition of legacy geodata

To follow the thread of this investigation, it is important to understand what we mean by legacy geodata or superseded geodata. We want to illustrate the concept with the example of the paper map. When printed, a map has an edition year. As an example, the scale 1:10,000 is published every five years and the scale 1:100,000 every 10 years. Imagine that both scales were published in 2010. In the year 2015, a new edition of the scale 1:10,000 is distributed but not the 1:100,000. The 2010 edition of the larger scale is superseded in 2015 but not the smaller scale, because there is no newer version of that data yet. We use the term superseded because the word 'historic' implies older data. Applying this principle to aerial photography, every new flight at the same resolution over the same area would supersede the previous flight. It is more complex to decide if data are superseded under changing capture criteria. Using digital terrain models (DTM) as an example, the technological advances of the capture instruments have evolved at such a speed that every new data capture is done in a higher resolution. It therefore depends on the use case if previous versions can be considered superseded. The participants of this study received the following explanation of superseded data:

> *[Superseded data include] all versions that are not the most current. I prefer to talk about superseded data rather than historic, because there is a certain expectation for historic data to be antique. This study covers all ages of superseded data: the data that just became superseded one week ago, as well as the data that is more than 100 years old.*

For databases, it becomes even more complex to decide what is superseded. As long as a database has no spatiotemporal dimension, every feature can be overwritten, merged with another, divided into various features or deleted without the possibility to restore the previous stage. In this case, there are never superseded data; everything is always current. Databases that incorporate a temporal dimension contain both current and superseded data objects. Superseded geodata can have sufficient value to be kept in the same information system together with the current data for several years. Therefore, only when the whole database becomes obsolete and the information is copied to a new system is the database as a whole superseded. Nevertheless, the creation of legacy data out of such dynamic databases can be forced by taking snapshots. In versioned information systems, legacy data accumulate and cohabit together with the updated data. For more details about versioning and ways to extract legacy data, see chapter 3.1.2.5.1.

As a synonym of the term superseded data we use legacy data, which will be the most used term for historic data in this thesis. The term 'obsolete' is used here only in the context of technology in disuse. In its extreme form, obsolete data cannot be read any more because knowledge about its structure has been lost and/or the hardware and software to open it are not available any more. The risk that data will become obsolete decreases when standards are used: the more widely a standard is adopted the lower the risk.

Archives and legislators make an effort to include digital objects into the preservation processes and their regulations so far designed for geographic information hold on paper. We think of geodata as enhancements of paper maps and tables. To not deny its parentage, we include geospatial paper information in our concept of geodata. Additionally it can be transformed easily into digital data such as through scanning of old maps and handled alike any other raster data.

1.2.5 Definition of the term archive

Libraries, archives, repositories, data warehouses and data producers can all take on the responsibility to preserve geodata on the long term. We will mention these institutions when their distinct perspective on geodata management and preservation is emphasized. Nevertheless, to simplify the text we use the term archive to refer to any institution with the intention to preserve geodata over several decades.

The term ‘archives’ is also used for an institution’s holdings: data and documents. To help the reader when referred to data and documents in archives we will use archival material, archival holdings or directly use the terms documents or data, except when we write about the stages such holdings go through as explained in chapter 2.2.2. To refer to archival material in the different stages we use: active archives, semi-active archives and definitive archives.

1.3 Aims and objectives

This thesis has two main goals:

The first goal is to map current and future users of legacy geodata to help archiving institutions define their designated communities and therefore move closer to alignment with the OAIS reference model. Specifically, we are interested in knowing what different user profiles exist and what their characteristics are.

Furthermore, we want to develop basic knowledge about the way customers find and reuse legacy geodata, so that potential archives can better design appraisal processes, and decide service levels and components of a geospatial archival system. This second goal builds on the first one; the user profiles developed to reach goal one will form the basis for the selection of specific users.

To reach these goals we have set the following objectives:

1. Detect legacy geodata users. This will serve as a first overview on user types on which the user profiles will build.
2. Identify user characteristics and needs which are relevant for appraisal and other preservation decisions. The characteristics will form the theoretical model of a user profile that is to be filled with values.
3. Define the values of these characteristics for all legacy geodata user groups to obtain the profiles. Objective one, two and three will answer research question one.
4. Determine which user types are relevant for a long-term digital geodata archive.
5. Provide a rich description of user behaviour and needs towards legacy geodata in Catalonia of all the relevant user groups to allow data triangulation.
6. Detect needs for change, amendment or adaption of existing appraisal guidelines for geodata given by the user statements and propose such adapted guidelines. Those adapted appraisal guidelines will help designing preservation processes for legacy geodata and can speed the implementation of these processes for those who adopt the guidelines.
7. Exemplify the application of the user profiles on a real preservation situation of a geodata producer. The application will assist in testing the profiles and their components. With the objectives four to seven we will achieve research goal two.

As a contribution to the knowledge base in legacy geodata preservation, this thesis offers well-defined designated communities of legacy geodata, as identified by the expert study. The resulting profiles should be adoptable by any institution archiving geospatial data that intends defining its designated community in accordance with the OAIS reference model. The execution of the research steps has been guided by the in-depth analysis of a map producer that is worried about the long-term preservation of and access to its data. The interrogation of its users, manifestations of possible current designated communities, allows the review and eventual validation of the method used for defining the general user profiles. As in any case study, the result is first and foremost useful to the studied institution.

Nevertheless, as a significant example, the proposed adapted appraisal guidelines should be transferable to other contexts where appraisal of geodata in the context of long-term preservation is needed. We consider these user-informed appraisal recommendations to be a second contribution of this thesis to the knowledge base.

1.4 Scope

The two main goals of this thesis are carried out on different levels. The first goal, to define user communities, has a broad scope. The characteristics of user communities will be drawn within the scope of western culture users (North America and Europe). Therefore, the participating experts will come from different backgrounds and represent at least four different countries. The second goal, to develop appraisal criteria, is tailored to the Catalan context and the Institute for Cartography and Geology in Catalonia, Spain (ICGC),[4] which will be the studied institution. For that reason, we choose to capture only the needs of user communities in Catalonia, to be compared with existing appraisal criteria and guidelines, to find out if these guidelines serve the ICGC and to which point the needs of the user communities might influence them.

It is important to limit the technological extent of this thesis because every type of geospatial content entails its own challenges. We exclusively analyse graphic geospatial data (aerial photographs, maps and topic layers that depend on them), because they present an extra challenge to preservation: thematic spatial data layers must be interpreted in context with a reference layer (Bos et al., 2010). Additionally, the focus lies on geodata generated by the studied institution to create topographic maps. We do not address exclusively textual data, because consolidated file format recommendations exist and different technical solutions are available and implemented for their preservation (Ball, 2010; Reich & Rosenthal, 2009). Nevertheless, where textual data is a by-product of the map-making process or is complementary information necessary to understand the core geodata, such as in technical documentation and measurement instructions, they are taken into account.

[4] Institut Cartogràfic i Geològic de Catalunya - ICGC

It is important to understand that we are studying the users of a medium-size map producer with a particular regional and legal context (Catalonia). As seen in personal interviews, user interests vary from one culture to another. Cultural differences can influence research interests, administrative practices, leisure activities and in consequence legacy geodata use. As a result of the combination of the technological scope and the focus on the studied institution's users, we only document use of geodata derived from map production. This study is not concerned with meteorological data or space data because these kinds of data are not produced by the institution that served as the study case.

As regards the choice of selection and appraisal guidelines for the comparison, we make a distinction between selection guidelines primarily intended to enhance access to collections and appraisal guidelines for data preservation. The library attached to the ICGC digitises large quantities of historic maps and many other institutions do so as well.[5] Digitisation projects usually are initiated to give access to the digital copies. For that purpose, the scanned documents are integrated in a digital library and therefore subject to digital preservation actions for the entire digital library, if one exists. Selection criteria are applied before the digitisation process and can, to some point, overlap with appraisal criteria of inherently digital maps. Nevertheless, we did not analyse priority lists and criteria for digitisation projects as they do not reflect the specific challenges of inherently digital material. We also do not analyse selection for collection building by way of metadata acquisition without corresponding data. Metadata-capture-only is an option for libraries that intend to enhance identification of and access to geodata, but does not include preservation by the library. Therefore, it is not an option for the ICGC. But, criteria lists for appraisal of satellite imagery have been analysed, because the characteristics of the files are similar to those of digital aerial photography.

Finally, this thesis analyses and enhances only existing appraisal criteria and does not answer all preservation challenges along the data life cycle. Preservation procedures related to or influenced by appraisal might be touched on, but not deconstructed in detail.

[5] A collection of links to digitization efforts can be found here: http://maphistory.info/projects.html

1.5 Methodology

Digital preservation is a new field and the methods proposed in scientific literature to define the designated community lack tested and documented results. Nevertheless, it is important to analyse the appropriateness of proposals by Giaretta (Giaretta, 2011) and Kärberg (Kärberg, 2014) for our purpose of assisting appraisal and selection. Giaretta defines the designated community by its knowledge base. He proposes to divide knowledge into modules. A module can be software, an object or a skill. The knowledge modules have dependencies and rules such that, for example, the knowledge of the module 'text editor' depends on the module 'basic computer skills'. All members of a designated community should have the same characteristics (modules). He says

> *if u is a DC [Designated Community], its profile, denoted by T(u), is the set of modules assumed to be known from that community.*

The OAIS reference model requires that the necessary information to fill the gap between the modules the data require and the modules available to the designated community must be stored in the representation information and preservation package information. These two concepts are information elements of the OAIS reference model that we are going to explore further in chapter 2.2.3.1. Giaretta's definition is intended to help determine the representation information and check its completeness periodically. The knowledge base of the designated community assists only partially in appraisal and selection. Service levels can be built on the knowledge base of the designated community, but we do not learn what users ask for, how they use it and why. Therefore, we keep Giaretta's idea of the modules, but expand it to properties other than knowledge and describe the user groups by this set of properties (see chapter 5.1.1).

As described by Kärberg, The National Archives of Estonia presents a way to define the designated community by web analytics. It includes measurement of three levels of properties: frequency of visit (to the digital archive), means used to find the archives webpage (link, search engine or organic) and length of stay. The combination of these properties results in 27 user profiles. Further analytics such as download statistics provide insight into what objects are most required, such that inferences about needs can be made. The intention of the Estonian project is to build a procedure that is repeatable and as automated as possible to monitor the designated community. Monitoring the designated community and checking understandability of holdings against it is a requirement of the OAIS reference model. It is not possible in our case to measure properties of the designated community by web analytics, because there is no established archive and not all historic resources are online. This means that if we were to rely only on web analytics, the question of 'what' the user requires would be biased and we might miss important related data sets. Nevertheless, we recognise the importance of knowing the significance of the user groups and add this as one of the properties that form the user profiles.

Giaretta's method is tailored to assist determination of the representation information and other metadata, and Kärberg's implementation is tied to a technique we cannot apply in our context. Furthermore, we could argue that web analytics gives us information about data requests but does not decode real needs. We consider that we need a broader approach to get insight into the 'who', 'how' and 'why' of legacy map use. In digital preservation, especially as it concerns the designated community, there is much still to explore. Indeed, when little is known about a situation or subject, qualitative methods help form the knowledge base on which further research will build. The range of qualitative methods used in this thesis spans interviews, questionnaires, focus groups and a Delphi study. Each method has its justification in the context of the evolution of this thesis and its scientific foundation, as explained in the following chapters.

1.5.1 Mixed methods

The expected results of the two research questions are of a different nature and justify a mixture of methods. The first question intends to formulate universally valid characteristics of yet-unknown legacy geodata users that in the future should be measurable and transferable. To answer the first research question, a basic research design was chosen that should contribute to fundamental knowledge about potential designated communities. We used the Delphi method, further defined in chapter 1.5.5, that includes human judgement; this method is known for its ability to predict and can deliver qualitative or quantitative results. The second research question strives to deepen our understanding of the users of our studied institution's legacy geodata, find out what data and services they need and how they react to selection and appraisal. The outcome of this goal is relevant primarily for the study case and drives at solving a problem. Therefore, a qualitative, explorative method was chosen to answer the second question.

Why not statistical methods? On the one hand, consolidated long-term digital archives of geospatial information are just emerging, so it is hardly possible to find current users to question about their interactions with the material. This means that few users are available, and the people who are willing to be interviewed or to participate in a survey are always only a part of those potentially available. On the other hand, the interpretation of user statistics will only be an option in the future if legal, technical and logistical conditions of a fully functional geodata archive allow it. So far, automated user statistics to give statistically significant answers are not available. In addition, they cannot give the necessary insight into the 'why' and 'how' that is desired to answer the research questions.

1.5.2 The influence of market research in this study

The promising possibilities of combining geodata with data of other fields of science, joined with the additional challenges for users to handle digital information, demand new archival services. The statistical and reporting techniques that were used in early user studies do not fit for the creation of new services, as nonexistent services cannot be empirically measured. Dent Goodman (Dent Goodman, 2011) asserts that 'librarians and others interested in discovering how users work and what they need can learn a lot from the practice and application of qualitative research in other areas'. For this study, we turned to market research for its long-standing practice focusing on user experience.

Market research can be performed at several stages in the implantation of a new product. At an early stage, it can help define the characteristics of the product or service. It can also evaluate user reaction to early design. In general, it explores the reaction of the market to the product and searches for a market segment – a user group that shares characteristics or needs.[6] A product or service is optimised for a market segment, as would be the archival service for the designated community. Nevertheless, market research has a short to medium time span, while digital archives must consider the long term. Several recent efforts bridge the gap between market research and the need for foresight over a longer term in an ever-

[6] Source: Buisnessdictionnary.com (http://www.businessdictionary.com/definition/market-segment.html)

changing society (Malanowski & Zweck, 2007; Postma, Alers, Terpstra, & Zuurbier, 2007). Following these examples, we used the Delphi technique to get a better insight into prediction of technological changes and the profiles of presumed user groups. Through a literature review it was confirmed that the Delphi technique is fit for academia (Landeta, 2006).

1.5.3 Case study with preservation audit

A first necessity was to understand the field of map production and its data. Therefore, a partner was chosen that showed interest in the thesis general subject and that could serve as a study case further on. Thanks to this partner the knowledge base for this research was acquired.

Because of the need for subsequent user interviews, the chosen institution's production had to have sufficient impact in the region to provide a suitable user base. The impact on other levels, such as the influence the institution has in the field, is also a reason to choose an institution for a case study (Patton, 2015). We studied the Institute for Cartography and Geology in Catalonia, Spain, because it offers service to users interested in antique maps through the map library and provides other legacy geodata free of cost on the ICGC's website. Part of its legacy data is distributed to public authorities and used over the internet by the general public equally. This meant that there was already a user base to study. All other geodata producers in Catalonia are much smaller, and they offer data to a restricted group of users that would be much more difficult to identify. Furthermore, the ICGC plays an important role in the development of competences in the field of cartography for other producers. Due to its size and importance in Catalonia, the distribution and extended use of its legacy data and its impact on other producers, the ICGC reaches significance as a phenomenon of interest.

The ICGC has a regular production, an affinity to research projects and is aware of the preservation problem (Anguita, Montaner, Oller, & Roset, 2012). Furthermore, the ICGC hosts the Catalonian map library inside its walls. The map library works as guardian of the map heritage of Catalonia, and in this way, connects two fields of science that at first might seem alien to each other: archival science and geosciences. The ICGC has a legal mission to create and preserve the map heritage of Catalonia[7] and therefore has an additional motivation to spur curational efforts (Montaner, 2008; Montaner & Capdevila Subirana, 2010). The expertise of the library staff and the interest of the technical department in preservation topics are a perfect symbiosis of expertise and make the ICGC an ideal partner for the project.

To get an idea of the current preservation situation at the ICGC we applied the Trustworthy Repositories Audit & Certification: Criteria and Checklist, a methodology usually used for auditing existing long-term repositories. Although preserving digital content responds first of all to a technological problem – the obsolescence of formats, hardware and software that can prevent future access to data – in reality its implementation is not only an IT issue. Answers to legal, financial and administrative questions regarding the system must be sought as well. For this reason, the methodologies for auditing computer systems, among which is ISO 27001:2005, are not sufficient because they focus on just one aspect of the problem.

In this context, several methodologies adapted to risk assessment and trustworthy digital repository audit have appeared and are being put into practice. These include:

- NESTOR Kriterienkatalog vertrauenswürdige digitale Langzeitarchive
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
- Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), now ISO 16363
- The Geoarchiving Self Assessment Tool

In 2007 the Network of Expertise in Long-Term Storage of Digital Resources (NESTOR) working group under the Ministry of Education and Research of Germany developed the *NESTOR Catalogue of Criteria for Trusted Digital Repositories* (Dobratz & Schoger, 2007); the second version, published in 2009, consists of a structured list of instructions a digital repository must comply with to be considered secure. The principle criteria are

- adequacy – acknowledgement that no absolute standards exist,
- measurability – all indicators have to be measurable,

[7] Ley 16/2005, de 27 de diciembre, de la información geográfica y del Instituto Cartográfico de Cataluña

- documentation – the objectives, specifications and implementation have to be documented, and
- transparency – part of the documentation has to be public to inspire trust in stakeholders.

Currently, NESTOR is the method of reference for auditing repositories for scientific publications and other content at German universities.

DRAMBORA (Digital Repository Audit Method Based On Risk Assessment) arose from a research project funded under the framework of the European Union and led by the University of Glasgow's Humanities Advanced Technology and Information Institute and the National Archives of the Netherlands (Quisbert, 2008). It takes the form of a checklist that allows for confirmation of the level of implementation of digital preservation policies in a repository. The scope of DRAMBORA is narrower than that of NESTOR, as evidenced by the fact that it is not intended for external audit and certification but focuses on internal audits.

TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist) started from studies initiated in 2005 by OCLC and the Center for Research Libraries of the United States (Hank, Tibbo, & Barnes, 2007). TRAC, published in 2007 (OCLC, 2007), is a complete method with a form similar to NESTOR, developed as an external audit mechanism. It has been tested in some of the main preservation repositories in the United States – Portico, HathiTrust and Chronopolis – (Center for Research Libraries, 2010, 2011, 2012). This experience led to a new version of TRAC under the name of TDR (Trusted Digital Repository Checklist), (Consultative Committee for Space Data Systems [CCSDS], 2011), which was approved as ISO 16363:2012 in 2012. It is expected that it will soon be possible to conduct certification audits under this standard.

The Geoarchiving Self Assessment Tool (Geospatial, 2010) is a checklist developed by the GeoMAPP[8] project partners. It is addressed to geoscience institutions that store their own data in order to enable them to objectively evaluate their archives' potential to preserve geospatial data or test their repositories current archiving practices. The criteria are classified in three categories from basic to advanced. Each of these levels covers the following areas: plan sponsorship and project governance, current programs, human

[8] GeoMAPP is an abbreviation for Geospatial Multistate Archive and Preservation Partnership (http://www.geomapp.net)

resource requirements, data requirements and technological requirements. The questions about these subjects are very detailed, but are not accompanied by indications on how to answer the questions, as TRAC does, for example. The Geoarchiving Self Assessment Tool is specialised for geodata, but tailored to the U.S. legal environment and meant to serve public archives. Because it is a recent product, there are no reports or experiences concerning the Geoarchiving Self Assessment Tool available yet.

TRAC was chosen to audit the preservation situation at the ICGC because it was already internationally recognised and a candidate to become an ISO standard (which it became in 2012). In addition, it was designed for external audits and included guidelines to the different criteria. TRAC is now known under its name *ISO 16363:2012 Space data and information transfer systems – Audit and certification of trustworthy digital repositories*.

ISO standard 16363:2012 is meant to evaluate the trustworthiness of repositories whose mission is to preserve digital objects. Currently the ICGC's mission statement does not implement long-term preservation as an objective and its information technology system is not managed with this purpose. Therefore, the norm cannot be strictly applied to the ICGC and was not used for certification. In this case, the standard was used to detect how well current processes at the ICGC conform to the OAIS reference model. It was necessary to adapt the auditing standard to better fit the technological and organisational context of the ICGC. Where components of an OAIS-conforming repository were missing, questions were omitted or terminology was adapted to similar, existing processes. All questions had to be translated into Spanish for fluent interviewing and to minimise misunderstanding.

The auditing process was carried out in three phases:

1. Initial contact between the ICGC and the auditors: In two meetings, technical vocabulary that was to be used in the auditing process was aligned, and the ICGC presented its administrative and technological structure.
2. Audit meetings: The audit questions were answered by the chief of the computing centre of the ICGC in two meetings that took place on 13 April and 23 May 2012. Due to the internal and informal aim of the exercise, the auditors did not require documentation to prove the veracity of the answers.

3. Subsequent clarification: Clarifying information was received by email from the general deputy director of technology and from the chief of the area of Information Technology.

Results can be found in chapter 4.2.

1.5.4 Questionnaire

We decided to shape the methodology around three user groups.

- Current users of historic data
- Current users of current data
- Future users of historic data

We intended to draw current users of historic data from the cartographic library at the ICGC, the Catalonian deposit library for maps, the cartographic libraries at the University of Girona (UDG) and the Autonomous University of Barcelona (UAB). Those four map libraries were chosen, because they have important map holdings and have special procedures and consideration for them, whereas at other libraries and archives maps are just one of many document types. All four institutions collaborated. In autumn 2013 they distributed a questionnaire to the users willing to participate, by email or on paper. The UAB had just sent their users an internal questionnaire and therefore limited the distribution to four email addresses of professors. The questionnaire can be found in Annex 1. The ICGC's library, the UAB and the University of Girona together sent 52 questionnaires by email to known map users. In all the institutions except the UAB, staffs were asked to inform users about the questionnaire and lay out copies. It is not possible to determine how many users were informed of the questionnaire personally. This method proved to be insufficient, because the response rate was low with only 13 answers and the type of questions did not allow in-depth analysis. Additionally, we were advised not to contact institutional users of current data because they were not poised for participation in a survey. This led to the decision that focus groups should be used for understanding current users of all types. Nevertheless, where appropriate, we used input from the respondents of the questionnaire to complement the user profiles.

Future users obviously could not be interviewed directly and were determined by a Delphi study, which is explained in the next chapter.

1.5.5 Delphi study: lay out user types and challenges and drivers to archiving

The Delphi method was first used by the Rand Corporation in the 1950s. It is a technique used when appropriate historic and technical data that could be used for a statistical prediction are missing. The goal of the Delphi method is to reach a consensus of opinion from a group of experts, who answer questions in subsequent rounds. After each round, they receive feedback and are asked to review or argue their opinions. This process is repeated until the differences of opinion become acceptably small for the contracting entity. A classical Delphi study is conducted blind, which means experts names are not revealed to each other. This technique allows 'access to the positive attributes of interacting groups (knowledge from a variety of sources, creative synthesis, etc.), while pre-empting their negative aspects (attributable to social, personal and political conflicts, etc.) (Fildes & Allen, 2011). The Delphi method was originally developed to gain quantitative results. Nevertheless, it is usually combined with open questions that can be analysed qualitatively. In the first round, open questions are often asked to extract the most important subjects that subsequent rounds will quantify. The Delphi method is particularly well suited to new research areas and exploratory studies (Okoli & Pawlowski, 2004).

For this study, two international expert groups were formed: group one was for defining the user profiles and group two was for predicting the influences of technological changes on archiving and accessing geodata. Nevertheless, group two also answered some questions about user groups. We used the Delphi study to give us some points of reference about the different potential future user groups and to provide us with the necessary awareness of technological evolution to compare it to the designated community's knowledge base. In a Delphi study, the quality of the experts is crucial (Landeta, 2006; Okoli & Pawlowski, 2004). For our study, all participants had to fulfil the selection criteria so that we could reach a homogeneous group with expert knowledge on the subject. For group one, we selected people who had direct work contact with users of geographic information or government information, or staff of major geographic information providers. For group two, we selected staff of geoinformation creators or people with technical background in data management or geographic information systems. Every expert was given the opportunity to rate his or her own answers, according to the perceived understanding of the specific subject. Furthermore, as there were questions about the user groups defined in the first round, we asked how well each expert knew each user group. These two ratings were later used to weight the answers. The experts filled out an online survey, created with LimeSurvey, which had open and closed questions. A pre-question helped define user groups of legacy geodata. These user groups were then used in the first round, where the experts were asked whether the study should consider other user profiles. The resulting propositions were grouped and resulted in three new user groups, which were added in the second round. The first round took place in November 2013 and the second round from February to March 2014. The online surveys contained an introduction that provided important definitions, a measure intended to avoid misunderstanding of the questions. It was stated several times, during recruiting and in the introduction to the online surveys, that the study would inquire about users of legacy data. The online surveys of the second round can be found in annex 2 and 3.

1.5.6 Focus groups: deepen knowledge about user needs and expectations

Because we cannot easily deduce current and future user needs from academic literature, primary contact with the user becomes more important and justifies a method that involves the user directly. Furthermore, the OAIS reference model suggests interacting with the archive consumer to monitor the designated community. Through direct contact we can find out why users access certain information but not others, which would not be possible by analysing automated user statistics. Understanding why a service is not used might be as interesting for a service provider as understanding why it is used. We searched for a technique that allowed us to hear experiences and register the users' thoughts and reactions related to the research topic.

Between the two options for personal interaction with the user – personal interviews and focus groups – the second was chosen. We think group discussions are particularly appropriate because they are also used in market research and applied research to guide program, policy or service development (Krueger & Casey, 2016). Focus groups enable small group discussions that provide qualitative data on a chosen topic. A facilitator asks the questions and ensures that the participants stay on track and do not get off topic. Focus groups have the advantage that in a single interview a variety of perspectives can be gathered, and participants might discuss ideas that would not have occurred to them on their own (Gorman & Clayton, 2005). Classical focus groups are conducted with 5-8 participants (Krueger & Casey, 2016; Patton, 2015). This study used smaller groups due to the difficulty of gathering several people in the same place at the same time who, in addition, had to share the characteristic of having used legacy geodata. The lower number of participants increased the time for each participant to share his or her thoughts and assisted the moderator in assigning turns if necessary.

The focus group facilitator was guided by an open interview protocol comprising three parts. The first part sought to understand use and use cases of legacy data by the participants. Each participant talked about his or her personal experience, which was intended also to make participants feel at ease with people they did not know. The purpose of the second part was to share and capture ideas for assessing needs. The third and final part of the focus group interviews was dedicated to gathering opinions and perceptions about defined appraisal strategies. In case of doubts about a participant's professional profile in the course of the interview he or she was asked with which user type denomination proposed by the Delphi experts he or she identified. The advantage of focus groups over personal interviews should be particularly evident in the second part, when interactions between participants can enrich ideas.

With the help of a question about similarity between user needs of different profiles, we clustered the user groups proposed by the Delphi experts into six clusters. During the focus group process, we interviewed members of four of the original six clusters. Two clusters were discarded for the following reasons: One cluster consisted of architects, engineers, construction workers, emergency response planning teams and military users, who were said by the Delphi experts to be interested in younger geodata and might never use a long-term data archive. Related to this cluster were institutions that create and maintain geodata. The study case belongs to this user group, and we do not intend to analyse its needs because it is much more sensitive to its own needs through its business processes than could be detected by a focus group. The other omitted cluster consisted of environmentalists and conservation agents of the non-natural environment. After analysing their characteristics from the Delphi study, we saw that the needs of these user groups would be satisfied with a service designed for the general public. The general public is part of one of the included clusters and was defined as a potential designated community. Of the remaining four user clusters – historians, geographers, urban planners and general public, we interviewed 35 members (26% female). All interactions with the focus groups were filmed, except for one, where only the sound was captured according to participants' agreement (Annex 5). The sound and film recordings were transcribed to text and analysed together with the written answers. More about the data analysis can be found in chapter 1.5.8.

All employed sampling methods used purposive sampling strategies and criterion-based case selection. The criterion that all participants had to fulfil was being a user of legacy geodata. In this sense, homogeneous sampling was pursued, although heterogeneity was the goal, to represent the variety of user groups. The recruiting strategies used reached from opportunity sampling to email invitations to the snowball technique, where one interested party recommends another. A workshop at the ICGC related to the topic of this thesis was organised. The opportunity was seized to invite the attendees to participate in a focus group. The direct contact increased the acceptance of subsequent interviews. Because geographic information is produced and used in the first place by public administration, many workshop participants came from this sector. For most public administration tasks, current geographic information is needed but some legal inquiries and sociological, economical or environmental decisions must be based on legacy data to gain solidity (Ariza López, 2012). That is why public administration staff are potential participants who fulfil the selection criterion.

The Delphi study indicated that another group of geodata users comes from scientific fields: academics and students of history, geography, economy and social sciences. Direct emailing based on a public list of members of relevant departments at all Catalonian universities took place. In the case of 42 institutions no member lists were publicly available and we used general contact email addresses or phone numbers. Members of 110 different institutions or departments were invited to participate in the focus groups. Finally, twelve users of the general public without known affiliation were contacted of which we recruited six. In total we contacted 1,087 people, institutions, departments or associations and groups specifically for the focus groups. Of the initial 44 people who responded positively 36 participated. Thirty one participants were interviewed face to face in groups of two to four participants each, except for one personal interview. Five participants answered the questions by email due to their incompatible schedule with the interviewer. Interviews with the remaining eight professionals did not take place due to their changed availability. The participants represented 16 different institutions or associations related to geodata or legacy data. Six participants did not have an affiliation to such a group or employer.

Each contact was asked to gather further legacy geodata users for the group interviews (snowball technique). Here a trade-off was made between a study design that would have preferred to avoid subjects of a hierarchical structure in the same focus group and the advantage to save time and coordination efforts, because only one interviewer was available. The disadvantage of possible domination of the group by a superior had to be accepted and was taken into account by the facilitator, who aimed to compensate when needed. Where it was possible, participants from different organisations or departments were mixed. The geographic variation of user provenance was respected; interview partners of all major regions in Catalonia – Barcelona, Gerona, Lerida and Tarragona – were included.

An effort was made to question members of all user groups in the cluster, which is why university departments of economics and major non-governmental organisations and associations were contacted to complete the profile of social scientists. Proposing participation to departments of law studies should have completed the profile of the policy makers and lawyers. Nevertheless, only a few members of economic departments and none of the additional groups answered favourably. The percentage of users of legacy geodata in certain areas is probably too low to find people willing to be interviewed. In Millea’s map library study, for example, no participant indicated legal inquiries as his or her map use purpose (Millea, 2005).

The interviews took place between July and September 2014. Written interviews were collected until October 2014. Most purposive sampling such as used in this research relies on theoretical saturation to determine the moment data collection can be stopped. Theoretical saturation is reached when additional interviews do not yield new information. The studies of Nielson and Landauer (1993) and Guest, Bunce, and Johnson (2006) intend to help determine the appropriate number of user interactions previous to data analysis. Nielson and Landauer (1993) calculate that 80% of usability themes are covered by six users in a homogeneous group. Guest et al. (2006) suggest that after 12 interviews in qualitative research no new themes will come up and that indicators of all general themes are present in the responses of six users. Indeed, transcriptions and analysis of interviews made in parallel with data collection for this thesis revealed repetition of answers and user attitudes towards legacy geodata. We reached a satisfactory level of saturation in the group of general public (9 participants), urban planners (9 participants) and geographers (13 participants). The group of historians with only five focus group participants did not reach the same level of saturation. Nevertheless, two participants in the initial questionnaire (see Annex 1) stated that they had a background in history and we could take into account their answers for the profile of historians. Additionally, the high effort put into recruiting and the low response rate in this group justified detaining the sampling process at this slightly lower level of saturation.

1.5.7 Additional data collection

In qualitative research, literally anything can be used as primary data for analysis. To answer the second research question, we needed a base of appraisal guidelines we could build our knowledge on. All appraisal guidelines discussed in this thesis directly related to geodata were extracted from the known geodata preservation projects, presented in chapter 3.1.1. Guidelines for research data selection were only included if they contained examples or further explications related to geodata or were developed by a geodata producer or geodata warehouses. Indeed, the view of scientific data repositories is important because digital geodata have scientific value, as discussed in chapter 2.1.6, and we do not want to miss criteria on the research value of geodata. S. P. Morris (2010) suggests that

> *Government records managers often appraise from the perspective of selecting for long-term retention those records that best document or capture the activities and information outputs of government agencies. In the case of geospatial data, it may be necessary to move beyond this approach in order to take into account the broad*

> *applicability of geospatial data, which contain valuable information of use in a variety of research areas that extend far beyond the intended use of the data.*

At first, retention schedules for geodata sets were also included, but analysis showed that they generally did not contain the appraisal criteria that were applied to them. Without this information, the retention schedules are not valuable for this research. The only exception is the retention schedule of the Kentucky Department of Library and Archives. It contains information about the reason for retention, which is why it has been kept for comparison. Because none of the appraisal guidelines have the status of being universally valid or at least of being a de facto standard, adding up criteria from all should lead to the most complete appraisal criteria list.

1.5.8 Data analysis

'A strong foundation for data analysis is solid data capture' (Patton, 2015). With the variety of material captured and the two research objectives in mind, we chose mixed analysis methods. On the one hand, results of the Delphi method could partly be analysed with statistical techniques. Open questions delve into some of the numeric results. On the other hand, the appraisal guidelines, focus group transcripts and some of the survey responses led to qualitative analysis. This combination of qualitative data with quantitative data is called triangulation and strengthens the study (Patton, 2015). The triangulation types used in this research are data triangulation and methodological triangulation as defined by Denzin (Denzin, 1978). Data triangulation refers to the use of various data sources while methodological triangulation refers to the use of multiple methods.

Slow change in user attitudes, needs and behaviour allowed data analysis to continue even after 2 or 3 years from data collection without the risk for the data to be outdated. Change in environmental factors such as new availability or accessibility of legacy geodata sets were taken into account and explained certain user preferences.

A relation of all the users who provided input through the different data collection methods can be found in Annex 6.

1.5.8.1 Data analysis of the Delphi study

For the analysis of the Delphi study results, the answers were exported to spread sheets. Mathematical calculations were applied where possible. Where averages were calculated, preference was given to the mean instead of the median to avoid having outliers bias the result.

It was not compulsory to answer every Delphi question, which is why some questions did not reach the minimum needed number of four answers. Many other Delphi studies recruit only four to five experts (Fildes & Allen, 2011), which is why we considered four answers to be sufficient. Likewise, if an answer was given without a sign of how well the expert knew the user group or if he or she indicated having no experience with that particular user group, it was considered 'not valid' and the answer was ignored in the calculation, except where otherwise mentioned. This procedure should enhance the reliability of the synthesis of all answers.

Because Delphi studies are iterating processes, between the first and second round a first analysis and synthetisation of the answers took place. Analysis and interpretation involved grouping of some proposed user groups and exclusion of others. Analysis methods used in the first round were repeated with the expert answers of the second round to be able to compare results between rounds. Nevertheless, some feedback led to new or adapted questions that hindered direct comparison but deepened insight into the specific aspect.

The user characteristics that were modelled through the Delphi study were extracted from different geodata preservation guidelines and general preservation handbooks. We used them to inform the modules that should build the user profiles. The ICGC took part in the elaboration of the Delphi questions as the technical complexity of geodata forced us to ask for technical characteristics too. The thorough input of its technical expertise led to well-defined questions intended to avoid misunderstanding by the experts. The user profiles presented in chapter 5.2.1.1 were formed by combining the characteristics (module values). For additional comparison of the user profiles a clustering method was used based on one independent Delphi question as explained in chapter 5.1.2.1. This question served as a control function. For the clustering of the proposed user groups the Louvain method for community detection (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) was used. Based on the similarity showed by the Louvain method and the other user characteristics, six user clusters were formed.

1.5.8.2 Qualitative analysis of collected written material about appraisal and user needs

There are a variety of approaches to qualitative research with roots in ethnography, philosophy, sociology, linguistics or humanistic psychology (Patton, 2015). Of the existing methods of qualitative data analysis, grounded theory might be the most famous. Grounded theory is the discovery of theory from data systematically obtained from social research (Glaser & Strauss, 1968). Grounded theory is an analysis method that has as its purpose the generation of theory (Trinidad, Carrero, & Soriano, 2006). The methods of constant comparison offered by grounded theory allows researchers to generate hypotheses after data collection, and after careful examination of the data (Auerbach & Silverstein, 2003). Coding of the textual data plays an important role in the process of reaching theoretical conclusions and goes through the following steps:

- Reading raw text
- Detecting relevant text
- Coding repeating ideas and finding their dimensions
- Grouping into themes (classification)
- Building theoretical constructs
- Finding theoretical narratives
- Formulating new research concerns

Because our goal is to build basic knowledge where there is none about the legacy geodata users, grounded theory seemed to be appropriate for our purpose.

The initial idea was to extract all arguments from the appraisal guidelines and all needs and demands from the user interviews using open codes. Open codes are developed during analysis of content, rather than having been determined in advance and limited to a certain list. Analysis started with the appraisal guidelines. All arguments found were used as codes. User statements related to those arguments were then marked with those same codes. Dissimilar ideas received new codes. By means of the codes, the appraisal criteria could be compared to the statements of the users and expanded or adapted. Coding was not completely free of influence: some codes stemmed from the modules of the user profiles, a theoretical model we constructed in advance. In the course of the analysis, modules became codes. Nevertheless, all statements were coded so that emergent ideas could be identified with open codes. The influence on the codes was not purposed, but naturally emerged from the research questions, which already guided the semi-structured interviews of the focus groups and therefore determined the comments and statements available for analysis as primary data. Therefore, it can be said that the codes were open and not biased. The iteration of analysis and refining codes corresponded to the proposed process in grounded theory. This type of analysis can be called comparative according to Strauss and Corbin (1998). By using this method, variations between the appraisal guidelines' values and the users' values were detected and variations in the pattern towards legacy geodata use were discerned.

For the coding and conceptual ordering of the raw data, the qualitative analysis tool Atlas.ti was used.

1.5.8.3 Remark on the linguistic dilemma

The author of this study is a German-speaking native, while the Delphi experts participated in English and the interviews were held with Spanish- and Catalan-speaking users. Having lived both in an English-speaking environment in the United States of America and Great Britain, and living currently and for several years around Spanish and Catalan speakers, we think the author has acquired the sensibility to capture the nuances of the spoken word in these languages. Nevertheless, when doubts arose, especially in written communication, the initiator of the expression was contacted to clarify, where possible.

To discuss the results many of the participants' statements are paraphrased. First, direct citations cannot be made, because the original expressions were in Spanish and Catalan. Second, direct citation increases the risk of identifying the participants, which would violate the code of ethics. When statements are interpreted, we refer to S1 to S13 for the thirteen participants of the questionnaire. G1 to G14 stand for the focus groups, while the individual participants are represented by P1 to P4, depending on the number of participants in one group. For example, the fourth participant in focus group 6 would be G6_P4.

1.6 Thesis structure

The previous chapters contain the introduction to this thesis that gave an overview of what will be achieved, how and to what purpose. In chapter 1.1 the motivation for this research is exposed. Chapter 1.2 defines the terms used in this thesis and their context. Chapter 1.3 and 1.4 address the research goals and scope respectively. The methodology is laid out in chapter 1.5 and this chapter 1.6 explains the thesis structure that will continue as follows.

This thesis is written with a background of information management. Therefore, all concepts used in areas of spatial information had to be acquired first. Chapter 2.1 reflects the resulting learning process and is addressed to readers with a similar level of prior knowledge. Readers with a background in the geosciences can skip chapter 2.1 but are encouraged to read chapter 2.2, a brief review of digital preservation and some archiving principles related to the subject of this thesis. Chapter 3.1 gives an overview of all actions that can be taken to promote preservation geodata and exposes the state of the art in digital geodata preservation. Chapter 3.2 dives deeper into the appraisal of legacy geodata and shows the results of the analysis of existing appraisal and selection guidelines. Section 4 presents the studied institution. It starts with chapter 4.1, a description of the ICGC, its geodata production, its library and its environment. Chapter 4.2 determines the ICGC's needs in terms of digital geodata preservation by means of a digital preservation audit. Section 5 presents and discusses the results. In chapter 5.1, a theoretical model of the user profiles that consists of several modules is presented and the opinions of the Delphi experts about users of legacy geodata are laid out. Chapter 5.2 discusses and answers the research questions. First, we combine the modules with the knowledge base about the users to build the profiles, which answer research question one (chapter 5.2.1). Second, based on the Catalonian user's needs for legacy geodata and their expectations and attitudes towards appraisal and future developments predicted by the Delphi experts, we discuss and propose adapted appraisal and preservation strategies, which answers research question two (chapter 5.2.2). Third, we apply the strategies to the ICGC and part of its production (chapter 5.2.3). Finally, in chapter 6 we reflect on the limitations of this research, evaluate the theoretical modules and methods, analyse the extent to which this case study and the Catalonian users' needs can be transferred to similar cases and draw conclusions on the improvement that could be made on the profiles by adapting the method.

2 Theoretical background

2.1 Theory of geodata

The terms geodata, spatial data or geospatial data are often used interchangeably. Explaining the meaning of geodata can be done by explaining the two concepts it inherits: geographical and data. Here is the definition used in this thesis:

The expression 'data' is used implicitly for digital data given that this thesis is embedded in the wider field of digital preservation. Most spatial data are not only 'born digital', which means that they existed first and only in digital form, but are also inherently digital, which means that they can be fully understood only when rendered by appropriate software and hardware. Data consist of numbers, text or symbols that represent raw facts (Longley et al., 2015). In the geosciences, sensor measurements, such as altitude, temperature and spectrum (of a pixel of a photograph), are examples of data. One category of sensor data is remote sensed imagery. It is called remote because of the distance to the captured object, such as a sensor placed in an aeroplane or satellite. Aerial photographs or images obtained with other technologies, such as radar or microwave, also pertain to this category. Data are assembled together in data sets and databases. In a narrower sense, information can be used as synonym for data. In a broader sense, information is distinct because it implies a certain degree of interpretation. Information is data that is communicated with a specific purpose. In the context of an archive, technological measures can preserve data but additional context must be captured to preserve information. Through the use of a geographic information system (GIS), raw geodata become information. We want to mention explicitly that the concept of data used in this thesis comprises raster images, such as scanned maps. Because they cannot be queried in the same way as numerical data they are sometimes not considered data. They are more commonly considered information, because of their incorporated representation.

Geospatial is an attribute used for 'datasets that have a geographic component, and that can typically be viewed as representing a portion of the Earth's surface in some way' (Janée, 2008). We do not include data representing outer space, nor spaces on such large scales that coordinate systems become irrelevant, such as the human body, although the concept ‘spatial’ includes both. Bühler affirms that ‘according to specialists, 75%–80% of all data has a spatial relationship to our living space and can therefore be called spatial data or geodata’ (Bühler, 2005), which would mean that there are a vast number of potential combinations of different data through their spatial relationships. Indeed, one of the distinctive characteristics of geodata is that some reference data can act as a framework to which other geographic information is matched. Geodetic data, coordinate systems, postal codes, topographic base maps and orthoimagery can all be used as such reference (Longley et al., 2015). Longley reinforces that adding value through linking multiple data sets is easy, as most often the coordinate system functions as the common linking key between data sets.

For an archive, the nature of geodata is an issue because raster images and individual textual/numerical data files are much easier to handle and preserve than complex databases. Databases are dynamic digital objects and can be preserved only when transformed into static objects such as when a snapshot is taken (Giaretta, 2011). Raster images are self-contained products and therefore easier to detach from the database without loss of functionality and understandability. Raster images are handled and preserved in many areas apart from geosciences, which is why experience and best practice are further evolved.

2.1.1 Purpose of geographic information

Geodata describe the world; they show what there is, where it is and how it is distributed in space. Geodata organised in a geographic information system can do more; they can explain how location-related processes work (Longley et al., 2015). The latter is necessary when having to predict the future and assess risk related to location.

Geographic information contributes to the solution of many problems. Three aspects distinguish the problems for which geodata can be useful: their scale or geographic detail and extent, their purpose and the time scale they must cover. Even though Longley distinguishes between practical problems, which often must be solved immediately, and those driven by human curiosity (such as research), he also points out that in both cases the same methods and techniques are used.

Here is an example of a local (large-scale) and practical purpose on a very precise and short time scale: one of the analysed users had to determine if a specific edifice was or was not at a determined location in a determined year. A completely different case is the example of archaeologists using aerial photographs and other remote sensed imagery to discover potential new sites (Aqdus, Hanson, & Drummond, 2012). The driving motivation in this case is human curiosity; it generates the need for larger time-scales even though the geographic resolution might be the same. Monitoring cultural heritage (Risbøl, Briese, Doneus, & Nesbakken, 2015) is done with a practical purpose over a longer time scale. The following example might need smaller scale geodata – by analysing photographs or querying vector maps one can find out, for example, land use in that area or the coverage of forest compared to crop land (Vogels, de Jong, Sterk, & Addink, 2017). Land-use analysis is one of the recurrently mentioned tasks of the interviewed users. The purpose of geodata reveals why certain data are collected and other data omitted. Perhaps the most famous example of a medical map is the one where John Snow depicts cholera outbreaks in Soho, London by combining object classes: the buildings where cholera cases were known and the water sources. This medical map helped people understand the connection between the affected water sources and the cholera cases and ended up guiding decision-making by the township in 1854. Medical mapping is now a teaching subject (Koch & Denike, 2004). Thematic mapping is popular also in other disciplines, but is not the only way to do spatial analysis. The distribution of events, objects or concepts in space can be queried without visual representation.

Location-based services help users find places of interest based on the position of the user. This is possible thanks to the Global Positioning System and the devices that capture position, such as mobile phones. Privacy is a major concern in this subfield (Yang, He, Sun, Lu, & Li, 2015). Current efforts are being made to improve accuracy, such as to use location-based services indoors as well. Location-based services depend on a map as reference.

By adding a temporal dimension to geodata, the evolution, fluctuation or course of an object in space can be observed. Some examples use time-stamped data over a short time range such as for traffic monitoring (Tang, Song, Miller, & Zhou, 2016). Others study long-term evolvement including data of several decades (Leyk, Boesch, & Weibel, 2006). The space-time information is extracted from a map or an aerial photograph and later visualized before a reference that provides context.

The internet and web-based geodata add new functionality to geodata – they promote integrative and shared decision-making. The additional advantages of web-based services allow for real-time space-time applications, such as for personalised public transportation information (Ma, Smith, & Zhou, 2016).

2.1.2 Structure of geodata

It has to be understood that spatial data cannot be equated with their representations. Numerical data can be represented in various ways as vector graphics or raster images. Vector graphics translate data into points, lines, symbols and shapes (Dingwall, Marciano, Moore, & Peters McLellan, 2005). Whether or not a geographic object is represented as a point of dimension zero or as an area depends much on the scale of the representation (Longley et al., 2015). In vector representations, data objects such as streets, rivers and forests take geometrical forms and might change shape when zoomed in. When the same data are rendered as raster graphics, the streets, rivers, etc. do not change shape, but lose resolution when zoomed in.

When thinking of geodata as structured information in a database, it is appropriate to talk of data objects with properties that define their state (such as altitude and date of construction) and methods defining the behaviour (what the object can do, such as connect to other objects). When data objects that stand for natural or artificial objects on earth are represented, they are often called features. Objects in an object-oriented database are organised in classes and are also called instances. In representations, classes correspond to layers. Images such as aerial photographs are also data objects that can compose a layer, but are not features.

In the following chapters the structure of geodata is explained, starting with the smallest component and gradually increasing in complexity.

2.1.2.1 Features and layers

The individual object represented in a map is called a feature. It can take the form of a point, a line or a polygon. The information we have about that feature is stored in attributes, for example the class of a feature: river, street, edifice, park, etc., or the direction of flows, or altitude for mountaintops, etc. Features of the same class must be of the same geometric type and are referred to as layers. The terms class and layer are sometimes used interchangeably (Longley et al., 2015). A typical example for a map layer is the transportation layer that could contain streets, highways, railways and other forms of transportation routes. Nevertheless, a layer does not necessarily consist of features; it could also consist of raster data. Raster data is not stored as a geometric object, instead it is stored as a continuing surface, such as in the case of aerial photographs or rasterized maps, but also any kind of remote sensed data with a grid of measure points. Often the combination of the two data types is useful such as for the display of non-topographic information (e.g., a thematic map showing the extent of an archaeological excavation) on a rasterized map or the vectorised occurrences of a biological variant on an aerial photograph.

2.1.2.2 Tiles and series

Tiles are individual parts of represented geodata of the same scale. Their existence is due to technological limits of representing a continuing space and has nothing to do with their intellectual content. Tiles are related to each other by the fact that once joined, they form a whole map of the same scale. To understand a tile, imagine a monographic printed map, limited by the size of the paper, that is part of a series of similar maps, which together cover a whole country or another area of interest. Tiles are used in a specific technological context when rendering or interacting with maps that should be smooth, and also when loading a whole map would be too slow. This means that with further technological progress, this technological solution might not be needed any more. As we said, tiles of the same characteristics form a series. Aerial photographs taken at the same resolution for the same purpose can also be considered a series. A time series is the same data collected at regular intervals. For example, all aerial photographs at the same resolution of the same area taken in different years make up a time series. For comparative research, time series are of special interest if the technical specifications of capture and semantics are homogenous or at least documented. Problems appear when capture rules or representation rules change from one version to the next, as happened recently with the 1:25,000 scale in Switzerland. In Figure 3 we see how altitude lines in the upper tiles are represented at a more frequent interval than in the lower tiles.

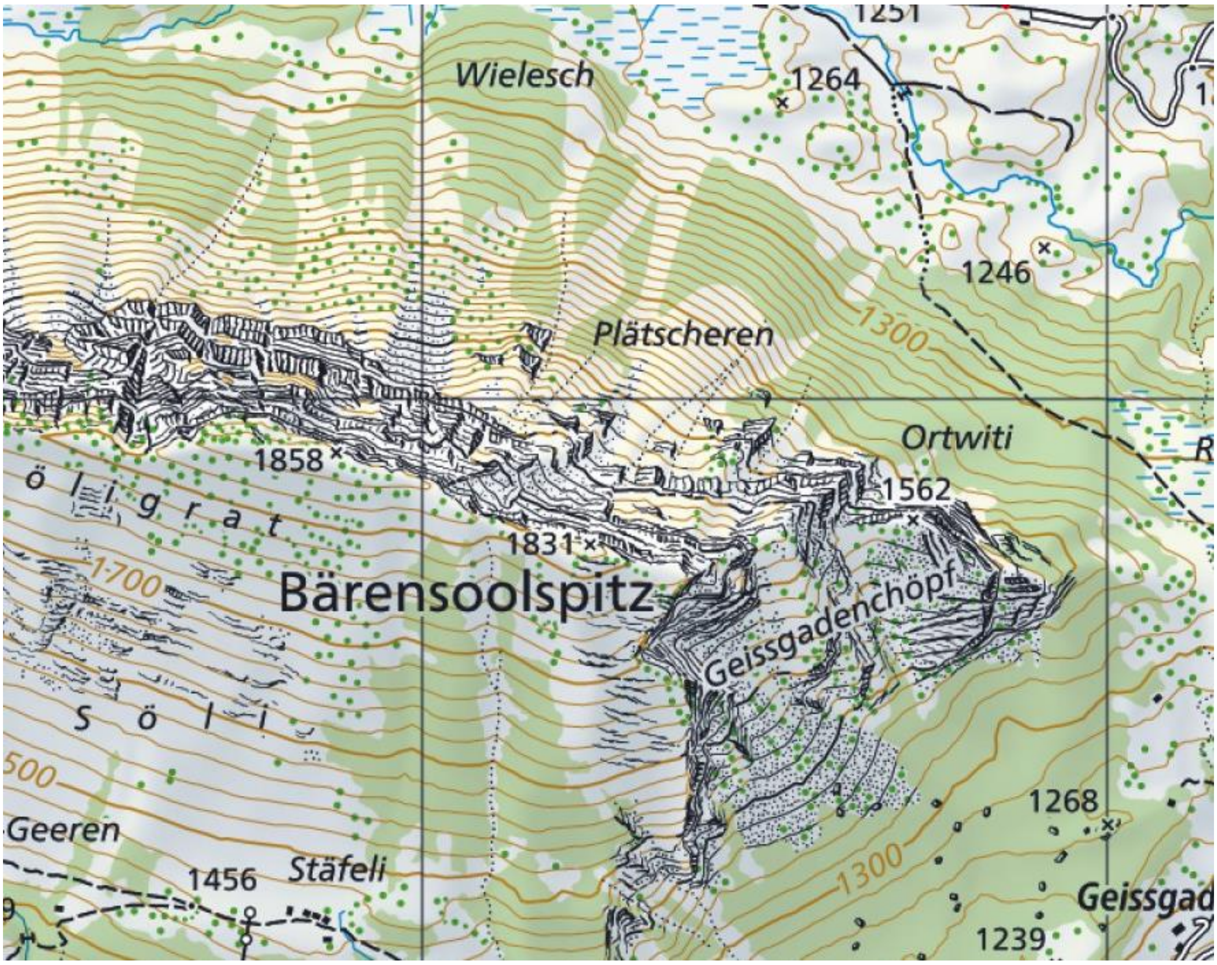

Figure 3: Extract of the map scale 1:25,000 in Switzerland. Tiles are delimited by grey rectangle lines. The upper three tiles show altitude lines at a double frequency than the lower tiles. Upper tiles pertain to more recent data capture and representation under new rules.

This is purely a question of representation and does not affect the data, but other rules, such as what density of trees is to be considered continued forest, are reflected in the data and can lead to significant change from one version to the next. Staying with the example of tree density, this would mean that in one year much more forest surface might be registered than in the next. At first sight, this could look like a significant loss in forest surface. Information about the capture rules and their changes is needed to interpret the data correctly.

2.1.2.3 A georeferenced file

A georeference specifies the location on earth of a data object. Through a georeference, distinct objects can be mapped together. Postal codes and city names can serve as georeference, although metric georeferences such as coordinate systems are more common, because they are more accurate and allow distances to be calculated (Longley et al., 2015). Any data object can have a georeference. An image such as an aerial photograph or an orthophoto in a JPG or TIFF format does not natively have a georeference, but for many purposes, it is useful to add one. It should be stored in a way it can be recognised by machine, connected to the image file and used immediately. This information might be embedded in the image file itself or in a separate file related to the main file. Georeferenced files that do not store the coordinates inside the file are exposed to the risk of losing their georeferences when handled. Indeed, any compound format risks losing the relations between its parts in transformation and migration operations (S. P. Morris & Tuttle, 2007). Not only raster images but also vector images and individual data objects can have georeferences. Like JPG and TIFF formats, a shapefile can be accompanied by a world file to define its geographical location. A shapefile contains vector data and is composed of various files in a file structure. The shapefile is a proprietary format invented by the *Environmental Systems Research Institute* (Esri). Even though its source code is not openly documented it has become a de facto standard for data exchange and many open-source viewers can render the file (Dorfey et al., 2009). Due to its widespread adoption and relative simplicity compared to geodatabase files, some institutions consider the shapefile format as a candidate for long-term preservation (GeoConnections & Hickling Arthurs Low Corporation, 2013; Geospatial Multistate Archive and Preservation Partnership [GeoMAPP], 2011b). Therefore, we are going to refer to shapefiles and use them as example in this thesis. World files are simple plain text files that contain the pixel size, rotation and x and y coordinates of a raster image. They are linked to the image file by maintaining the filename and amending the file extension by a ‘w’. For example, the file extension of a JPG file would be jpgw or jgw.

2.1.2.4 The geographic database

Geodata are more useful when stored in a database where they can be connected to all the other features, layers and textual or graphic information. ‘A data base is an integrated set of data on a particular subject’ (Longley et al., 2015, p. 194). Databases can be managed by database management systems that organise related data into object classes, tables and series. The database allows querying geodata in many ways that would not be possible when the information is isolated. To provide all functionality, the database might use its own proprietary file formats. However, when geodata need to be transferred they are often exported into a simpler format. While some information is compulsory to data interpretation, other information might not be included in the selected export format. When data are extracted from a database and transferred to a new format, they often lose functionality.

The information that software needs to render a file is partly given by the software itself and partly contained in the file. Some file formats depend heavily on information stored outside the file, for example in the software or in a separate file, so that they can be rendered as they were created. Properties related to representation of data might not be stored in the file, but tied to the software. For example, the colour of a layer or the thickness of a line is not necessarily memorised when data are extracted from the database and an Esri shapefile is created. When the representation properties are important they are stored in a separate file, which might be lost when transferred or handled. Depending on the purpose of the data, the colour or thickness of a line can be significant to the user. In this case, we talk about the significant properties of the data. Digital preservation is expected to maintain significant properties of data. This is critical, especially when data must be transformed into new structures.

When a geographic database works together with a visualization system it is called a geographic information system, abbreviated GIS. All large geographic information system applications include database management system software technology. Databases are updated to provide current information when features on the covered ground change. Updating might affect only small parts of features or layers and occurs in irregular intervals. Maps rendered directly from the geodatabase will always show the most current update, which makes them more useful to many applications than printed maps or generated raster maps where updates show with delay. Most online rendering tools, though, still rely on raster images for display.

2.1.3 Versioning: the temporal dimension of a geographic information system

Many geosciences institutions use geographic information systems (GIS) to visualize data and maintain them up to date. These systems assist in data management and dissemination. Because of the need for updated information, old versions in GIS are generally overwritten by data producers (Bethune, Lazorchak, & Nagy, 2009; Dingwall et al., 2005; Geospatial Multistate Archive and Preservation Partnership [GeoMAPP] & Library of Congress, 2010; S. P. Morris, Nagy, & Tuttle, 2010; S. P. Morris, Tuttle, & Essic, 2009). In geoscience, new capture can be planned on regular bases or triggered by singular events such as new construction or a natural disaster. The risk of being overwritten is higher for vector data layers than for remote sensed data (S. P. Morris et al., 2009). Substantial progress has been made to incorporate the fourth dimension of time into GIS and permit advanced versioning (Yuan, 2008). In databases that allow the capture of this fourth dimension, superseded and current data cohabit.

There are three approaches to modelling temporal changes in time (see Figure 4 below): the traditional snapshot database, the object-oriented approach and the event-oriented approach (Gantner, Waldvogel, Meile, & Laube, 2013). Snapshots are copies of the database at a certain moment and show the data as they were at that moment in time. Various snapshots of a database can form a time series. The snapshot model is built on the relational database and captures change only implicitly by repeated storage of entire data sets (first line of Figure 4). To detect change, different stages (snapshots) have to be compared to each other (Worboys, 2005). Each new capture supersedes the previous one, even though the administrative validity of the data content that did not undergo change is still intact. The object-oriented approach, as its name implies, is based on the object database. Each change in the object model generates new versions of the affected objects that can be placed in the hierarchy of time. As seen in the second line of Figure 4, only the objects that were affected by the update are captured again. Such databases allow time stamps for objects or attributes. The object can adopt stages of validity – active or inactive – and its attributes store temporal data such as date of creation, date of modification and date of inactivation (elimination or fusion into a new feature). The two basic possibilities for versioning objects are: either the attributes have their own time stamps and only attributes affected by change are duplicated, or the object as a whole uses a time stamp. In this second option, when an object is affected by change a new object with a unique identifier is generated (Lohfink, Carnduff, Thomas, & Ware, 2007). The superseded data in the object database comprise only the attributes or objects that had been affected by change and have a new version even though the database itself might still be administratively valid. Objects and attributes with time stamps allow restoring all previous stages of a database. Finally, the event-oriented approach in its simplest form 'consists of a base map that depicts the initial state, which is then amended by subsequent changes recorded in a transaction log' (Gantner et al., 2013). Line 3 of Figure 4 shows that only what is amended to the red object is captured. From an archival perspective, there are advantages and disadvantages. On the one hand, because only the parts of features that are affected by change are stored, redundancy is minimal. On the other hand, it is hard to determine when data are really superseded. On a visual representation, a feature can only be considered superseded when it is completely hidden by the amended change.

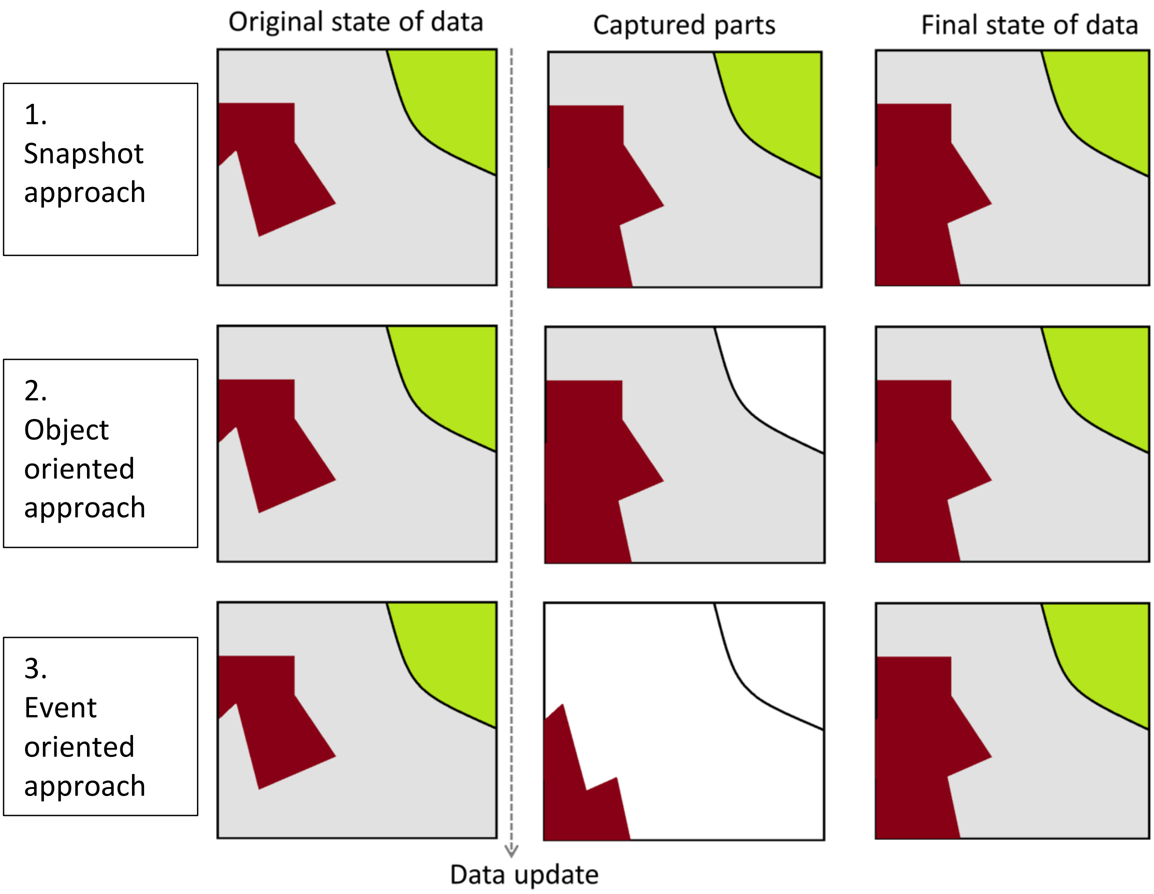

Figure 4: The three ways of capturing the temporal dimension of geographic data: the snapshot, the object oriented and the event oriented approach.

2.1.4 Difference between maps and geodata: functions and significant properties

Some geographic data take the form of a map. The definition of a map by the Cambridge Dictionary is ‘a drawing of the earth's surface, or part of that surface, showing the shape and position of different countries, political borders, natural features such as rivers and mountains, and artificial features such as roads and buildings’.[9] Therefore, only when geodata is graphically represented is it called a map. So-called base maps transmit the location and shape of basic natural and artificial features. The information they carry comes from direct observation or measurement of the terrestrial surface. This definition is given by the Catalan law 16/2005 of December 27 2005 (Catalonia, 2006). The base map is catalogued in the official registry of geographic information, and Catalonian government entities are obliged to use this official map data for planning and transmission of spatially represented information. Derived topographic maps, on the other hand, stem from base maps but pass through edition or generalisation. The process to render an aesthetic and readable map out of too-concentrated and abundant data is called cartographic generalisation. Generalising a map involves automated algorithmic processes but also human correcting and editing. Thematic maps represent information about non-geographic topics, such as those pertaining to biology, sociology, geology and many others, in space. To make it easier to read a thematic map, it is often overlaid on a base map. A map can be published on paper, electronically or not at all. Maps are made for a specific purpose and ought to serve this purpose primarily. When used in other contexts the information they carry can be hard to adjust or can even lead to misinterpretation (Hoebelheinrich & Banning, 2008). A map is conceived as the final step of processing geographic information, which makes it become an eligible candidate for digital long-term preservation.

A map has two technical properties that add to its complexity: the scale and the projection. Scale and projection only affect geodata when they are rendered on a two-dimensional

[9] http://dictionary.cambridge.org/dictionary/english/map, accessed on March 2017

surface. These two concepts are explained as follows: the scale is defined by the ratio the symbols on a map have compared to the real objects they represent. On a printed map this is a fixed ratio. Its legitimacy decreases with digital maps. A raster map changes the scale of its representation on screen when zoomed into the graphic, although in this case it loses resolution. A vector map increases its scale when zoomed in, and it maintains the resolution. Therefore, a specific scale is only true for a static view of a digital map. The projection entails distortion in at least one of the geographic attributes – direction, angle or surface – due to representation of curved objects on a flat space. Distortion also affects photographs taken of earth. Therefore, raw aerial photographs go through a process of correction to be of direct use to map production. The final photograph of earth is then called an orthophoto or occasionally an orthophotomap. There are many different geometric projections intended to serve the purpose of a map, maintaining one or the other of the map properties.

The queries that we can perform on a modern geographic information system that holds geodata are only possible because of the structured storage of the data. Even though it might look like some queries such as the extraction of a digital elevation model are executed on the map, they are performed by the GIS and answered by the geodata. The structure of geodata makes queries possible that are not imaginable with maps. This makes it so valuable to maintain the structure of the data when archiving them, because they might be queried in the future. Nevertheless, representations can be more practical for such purposes as sharing information that just needs consultation and avoiding complexity for the receiver, for example.

2.1.5 Value of legacy geodata

While some primary users replace superseded geodata with new versions, other user groups are interested in legacy data. The longer the time span a series covers, the more possibilities for analysis it offers (Conway et al., 2013). The main argument for long-term preservation in the geosciences – longitudinal research, mentioned by several authors (Moran, Hutchinson, Marsh, McClaran, & Olsson, 2009; Sweetkind et al., 2006) – becomes evident upon searching geosciences journals, especially in environmental science. Climate change and other kinds of environmental change analysis are among the many subjects that benefit from long-term data preservation (Beruti, Conway, Forcada, Giaretta, & Albani, 2010; Committee on the

Preservation of Geoscience Data and Collections et al., 2002; Erwin & Sweetkind-Singer, 2009; Erwin, Sweetkind-Singer, & Larsgaard, 2009; Harris, 2001; Janée, 2008; Shaon et al., 2011). Understanding change on our planet can help predict natural events, even prevent catastrophes (Beruti et al., 2010) and help governments to better administer natural and human resources and manage planning.

Some libraries (Erwin et al., 2009; Geospatial Multistate Archive and Preservation Partnership [GeoMAPP] & Library of Congress, 2010) have recognised the long-term value of geodata to their users and have engaged in preservation projects. One of the first challenges for these libraries was the selection of appropriate data. Part of the criteria given by the funding bodies for their preservation projects was that the information selected should be at risk and impossible or very difficult to recreate (Cruse & Sandore, 2009; Lazorchak et al., 2008). Because change is unavoidable, geodata collected at a certain moment in time cannot be re-collected at a later time. In one case it was argued that almost all digital geospatial data are at risk due to the sheer amount of data produced (Sweetkind et al., 2006), making it impossible for one institution to hold. Data can also be at risk by being produced or stored by small institutions or a single person, or ultimately because they are stored only in one geographical place (Erwin & Sweetkind-Singer, 2009).

As geodata producers, ESA, NOAA and NASA are conscious of the value of legacy data. ESA perceives (Beruti et al., 2010) and Morris confirms (S. Morris, 2013) a rising demand for time series which is why ESA recommends that its members archive geodata.

2.1.6 Is geodata research data?

At this point we would like to analyse whether geodata is research data, because this plays a role in the law on archives and document management in Catalonia and for the selection of appraisal guidelines. Tjalsma & Rombouts (Tjalsma & Rombouts, 2011) state that there is no precise definition of research data. They argue that 'research data' includes all research output other than publications, articles and emails created in a research environment. Research data is kept, among other reasons, for its value for new research that combines and compares data in new context (Palmer et al., 2011). Research activity at universities or at company internal research departments might produce thematic map data that are considered research data under this definition. Already archived research data is subject to new research input. Even though geodata creators such as the ICGC might have research departments, the data generated for map production do not come from research activity and are therefore not research output. Nevertheless, geodata often serves research. Their reuse possibilities as research input gives the data values that might not be considered in classic appraisal criteria but are considered in criteria designed for research data. Given that the Tjalsma & Rombouts' definition does not explicitly exclude research input from research data, and considering the data life cycle as a closed cycle of use and reuse, the data we focus on in this thesis (base map and intermediate data that served for base map production) might be research data.

2.2 Theory of digital curation and preservation

To understand the preservation of geodata, we would like to give a brief introduction to digital preservation in general. Preservation is more than conservation and storage. Conservation is passive archiving under conditions that favour physical maintenance of the object, while preservation includes actions that must be taken to maintain the readability and accessibility of the information. Both preservation and conservation are associated with a long-term commitment. The term storage, when used in the field of informatics, refers to saving information on an electronic medium. Although it has no long-term connotation, it is a condition for preservation. The term archiving is used by the fields of informatics, information science and geoscience, sometimes with a long-term connotation and sometimes without. For clarity, the term preservation or curation will be employed when long-term access must be provided. Preservation can be understood as the actions taken by the archivist once a record has come into his or her custody. Nevertheless, data quality depends on the treatment of the record during its life cycle. Actions taken before a record reaches the final archives can highly favour preservation. When favourable measures are taken before final archiving we will speak of digital curation, to include the entire data life cycle. For the purpose of this document, digital preservation refers to the treatment applied by the final archive.

Digitisation is one way of creating digital content, but digitisation is not digital preservation. Nevertheless, we often find the two terms used interchangeably. The concept of 'digitisation' refers to making a digital copy of an analogue object. It is a preservation measure for the analogue original and not for the digital copy. Obviously, the digital copy must be stored, but the storage space does not necessary undergo preservation management.

Even though we can say that digital preservation is relatively new, it has gone through a terminological shift that reflects the focus of this investigation. In the beginning, digital preservation was concerned with the more technological aspects of maintaining digital material readable. Now it is also trying to resolve organizational challenges of giving 'sustainable access' or 'long-term access'. Preservation and access are mentioned in the same breath by newer legislation in Catalonia, for example the law about archives and documentation and the law on legal deposit that we will present in the next chapter.

2.2.1 Data life cycle

Data management takes care of data all through the life cycle from creation to use and transformation. Good data management at the early stage of planning data creation includes planning its preservation (Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA, National Research Council, Division on Earth and Life Studies, & Board on Atmospheric Sciences and Climate, 2007). Planning the funds for preservation actions at the data creation stage has proven necessary to assure sustainable preservation beyond the primary life-span of the data.

From creation to reuse, data go through various stages, as shown in Figure 5. Data management makes sure all the steps in the data life cycle are fulfilled under the best conditions. The producer creates and ideally appraises the data before they are ingested into an archive. The archive performs preservation actions to ensure data reliability, usability and authenticity, such as defining preservation metadata and appropriate data structure and file formats. In doing so, it is maintaining data integrity. Long-term storage is the duty of the archive, while access can be a shared responsibility with an access provider. Ideally, the user locates, finds and reuses the data. For reuse, most data are transformed, for example by combining with other data, by sampling a subset or just by migrating the format. Transformed data can become a new archival object and the subject of repeated appraisal and preservation.

The following graphic representation of the data life cycle is inspired by the DCC curation lifecycle model.

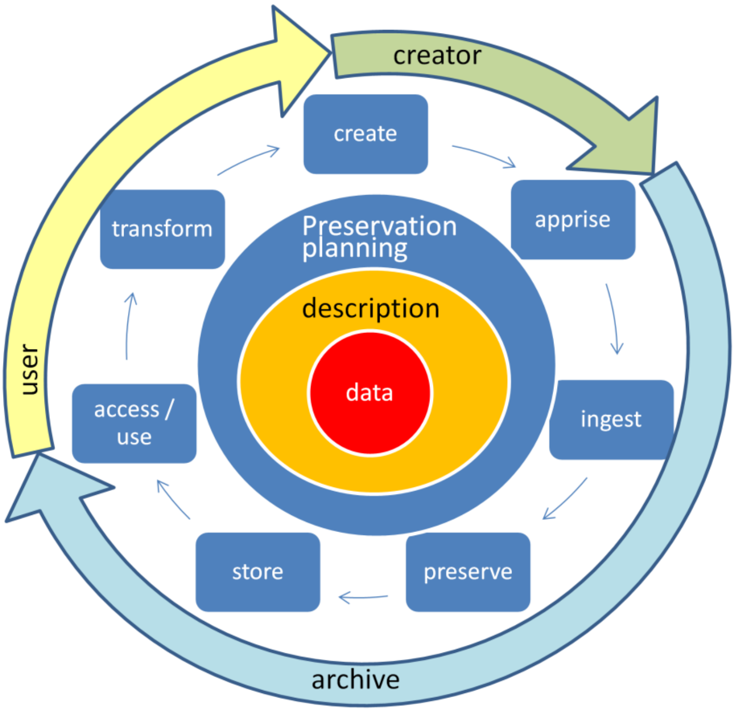

Figure 5: Data life cycle inspired by the model proposed by the Digital Curation Coalition.

At the centre of the data life cycle are the data, which are the objects of all preservation efforts. On the next level is the description of the data that is needed in order for them to be correctly understood and preserved. Preservation planning must assure all necessary actions are taken and processes regularly revised and updated.

The data life cycle model can be applied to geodata from map production. Due to the complexity of geodata structure, transformation is not only made by the data user but also can be a preservation measure. Data might be transformed to a standardised file format at ingest to sustainably store them (Projektteam Ellipse, 2013; Rönsdorf et al., 2013) or migrated at a later point in the preservation phase due to risk of obsolescence of the file format.

2.2.2 Active, semi-active and definitive archives

Documents do not pass directly from being used daily to being permanently archived. Neither do data. When data are current, they correspond to what is called an active archive. English-speaking countries use the term records for active archives. Active archives consist of documents that are susceptible to long-term preservation in the future, or are already destined for archiving by law, but are necessary for the daily tasks of the administration that created them. The law on archives and documents defines such documents as follows: administrative documents that a unit manages or uses regularly during its activities. The archives pass to the semi-active stage when use decreases or at a regular interval, as determined by a preservation plan. Catalan law defines these archives as consisting of administrative documents that, having concluded the ordinary purpose, are not regularly used anymore by the unit that created them. Before passing from the active to the semi-active stage, ideally the document collection will undergo 'thinning', which means elimination of copies or redundancies. At this time, a first evaluation for long-term preservation can also take place. Another evaluation will occur before the collection passes to the definitive stage. Because both evaluations, between the active and semi-active stage and between the semi-active and definitive stage, have similar purposes, Project Ellipse suggests relating the two appraisal processes. Values and characteristics of documents reported at the first step should form input for the second appraisal (Projektteam Ellipse, 2013).

Archives in the semi-active stage usually stay in the custody of the creator but are removed from their prominent and costly locations. The semi-active stage can last for several decades. Swiss legislation calls this phase, In the digital context, 'long-term availability' (Bos et al., 2010). The archives might be kept for legal reasons or they might still have occasional administrative or informational value. After the archives lose all legal and administrative value, they are evaluated for their historic, scientific and informational value. Archives with no long-term value are destroyed. At this point, the archives with value are usually transferred to their long-term storage place and are called definitive, historic or inactive archives. In Catalonian legislation, these archives are defined as administrative documentation that, having concluded their current immediate administrative effect,[10] possess primordial values of cultural or informative character. In transition between the semi-active and the definitive stage, archives go through a (second) cleaning process which might consist of selecting a representative sample.

With the advent of the digital age the linearity between the active, semi-active and definitive archival stages, which often entailed physical relocation, has become criticised. A more integrative concept of records continuum was invented that reflects how management of active and definitive archives are interwoven (McKemmish, 1997). Notably, in the case of continuously updated databases or management systems, the historic unit is inseparable from the administration of the whole. This means that new triggers for selection and appraisal must be found, and that systems designed for current management might need new security and trust features, because data stays in the system longer.

2.2.3 Importance of preservation standards – OAIS, TDR and PREMIS

The essence of archival work is to build trust. On the one hand, when a historian uses documents from an archive he or she assumes that they are an unchanged, original and truthful manifestation of what happened in the past. Every scientist that reuses research data from a repository must have trust in the digital object. So do the technical, legal and administrative staff that use data for reference. On the other hand, people or institutions that leave documents in the custody of an archive want to be sure that no unauthorised use will be made and that the documents or data will stay readable and unchanged over time. This stands in contradiction with the fragility of digital information; a single wrong click can delete or corrupt it. Trust is also involved in shared archiving, when one institution depends on others to store sufficient copies of its data (Berman, Kozbial, Mcdonald, & Schottlaender, 2008). They trust each other to maintain the storage environment as would be necessary for their own data.

The field of digital preservation has introduced several conceptual and technical solutions in its short life to address problems related to the digital nature of its holdings. Combined

[10] Translated from: vigencia administrativa inmediata

efforts to synthesise experiences led to the conceptual framework of the Open Archival Information System[11] reference model, which became ISO standard 14721. OAIS may be the most respected standard in digital preservation and many archival projects and technologies seek to comply with it. Archives that comply with standards inspire trust, because standards stand for quality and good management. Additional trust is given when the implementation is confirmed by an official institution. To check the conformity of a repository with the open archival information system reference model, an audit with ISO standard 16363 can be performed.

On a more technical level, there are the strategies formalised in the PREMIS standard (Preservation Metadata: Implementation Strategies). PREMIS addresses data management at ingest and inside the long-term repository.

Using standards forges interoperability between systems and data sets. Interoperability is necessary when passing on information from one archival system to another implementation. This could happen because of institutional fusions, reorganisation or cessation of operations of the original holder. Information might also be passed on to the next generation of archival systems in the long term. Institutions that follow the same standards will encounter similar challenges. This can lead to collaboration or cooperation in finding solutions and can be a way of saving resources.

We will briefly present the OAIS reference model, the TDR checklist and the PREMIS standard.

2.2.3.1 OAIS (ISO 14721:2012)

The OAIS reference model was introduced in 2002 as a recommendation by the Consultative Committee for Space Data Systems. It was accepted as an ISO standard in 2003 and was revised in 2012. It is conceived as a framework for a long-term digital preservation system and does not imply specific implementation. Therefore, it can be used for planning and designing an OAIS-conforming preservation system, but does not give concrete steps to take. The reference model guides management of the parts of the data lifecycle from ingest to access. The target of preservation efforts is the content information. Around the content information, a whole system of metadata and functions must be built to keep the content

[11] Consultative Committee for Space Data Systems (CCSDS). (2002).

understandable by a designated community. In digital preservation, content and certain metadata need to be tied together to maintain understandability and usability of the content. The OAIS model does not compel specific metadata elements but imposes types of information about a digital object that must be present in a long-term archive. These information types have many names, depending on what information they contain: preservation description information, packaging information, descriptive information and representation information.

Metadata directly related to the content is called preservation description information and contains information about the provenance, context of creation, reference or identifiers, fixity and access rights. The preservation description information identifies the purpose and context of creation, the producer and further holding institutions of the data, eventual relations with other data objects, identifiers unique to the data object that distinguishes it from other content such as a digital object identifier, a protective shield such as a checksum documented in the fixity information and access rights and restrictions.

Packaging information binds the preservation description information to the stored content by actual or logical relations. The packaging information identifies the components of the package and gives information about their size and location, such as in a file directory.

The descriptive information describes the content of a package in order to find, order and receive it. It can consist of only a descriptive file name or of a set of metadata elements that help to discover relevant content through a catalogue such as Dublin Core. The relation of the content, preservation description information, packaging information and the descriptive information of a package is shown in Figure 6.

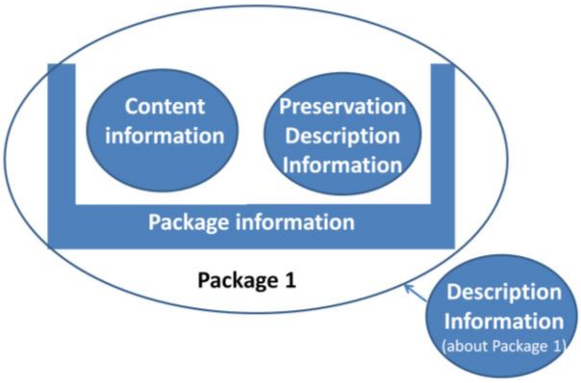

Figure 6: Components of a preservation package and its descriptive information. Source: Consultative Committee for Space Data Systems (CCSDS), 2002.

The representation information is responsible for appropriate rendering of the data to humans and computers in the future. This information type contains the description of the structural and semantic composition of data. It includes technical descriptions of file formats and digital objects such as applications that are necessary to render or run the data. The quantity and type of representation information depends on the knowledge base of the designated community. The OAIS must understand the knowledge base of its designated community to understand the minimum representation information that must be maintained. Examples of structural needs are file type descriptions, software requirements and platform dependencies. Semantic information helps future users to understand the data and properly interpret forms, figures and text. In metadata standards intended purely for content description, representation information is missing. To compensate for the lack of structural information in metadata standards intended for retrieval, a link to a registry where such information is available, can be established (Shaon & Woolf, 2011). An example of such a registry is the Library of Congress Sustainability of Digital Formats website.

The six main functional units of an OAIS-conforming repository – preservation planning, administration, ingest, data management, storage and access – are hereafter briefly explained.

An OAIS-conforming archive conducts preservation planning that embraces monitoring functions and determines preservation strategies, standards, package design and migration. The monitoring functions refer to the constant obligation for being aware of changes in the context of the OAIS, for example in respect to the designated community. Preservation strategies and standards should assess risk to the system. Therefore, they receive input from the other monitoring functions and recommend changes in strategies and standards to the administrator. Package design and migration is the function that designs new package models if necessary and tests migration options in detail. The technological or logical solution should be documented in the package design.

The administration functional entity is involved in system configuration, access control, archival information updates, audit submission, customer service and negotiation of submission agreements with the customer. The actual functions dealing with the content are ingest, data management, storage and access. Ingest handles the reception of submissions, quality assurance, generation of the archival information package and the related descriptive information and finally the coordination of updates of the database when the data are correctly stored in the archive. Data management is responsible for receiving updates to the database, its administration, performing queries and generating reports. The archival storage functions are receiving and providing data, error checking, managing the storage hierarchy, replacing media if necessary and disaster recovery. Finally, the access functional entity coordinates access activities, generates dissemination information packages and delivers responses to queries, orders and assistance requests.

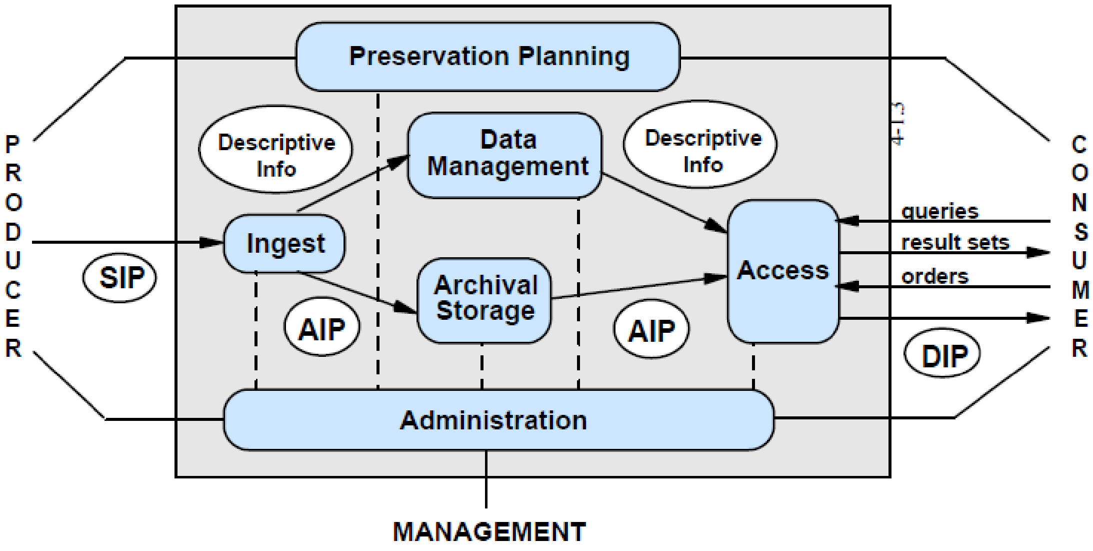

Figure 7: The open archival information system reference model's functional entities. Source: Consultative Committee for Space Data Systems (CCSDS), 2012.

The future standard ISO19165 should become an application of the OAIS reference model to geographic information.

2.2.3.2 TDR (ISO 16363; previously called TRAC)

The Trusted Digital Repository Checklist (TDR) is an audit checklist developed in response to the need to certify repositories that claim to be compliant with the OAIS reference model. Its intent is to provide objective measurement of trustworthiness by checking on the presence of the components of an OAIS-conforming repository. It shares terminology with and is designed to reflect the structure of ISO 14721. It is meant to be a checklist for external audit but can serve as support to repository system design. All certification criteria are accompanied by a list of possible evidence that could support the claim to meet the metric. The metrics are empirically derived and, as a whole, allow judgement of 'the overall suitability of a repository to be trusted to provide a preservation environment that is consistent with the goals of the OAIS' (Consultative Committee for Space Data Systems [CCSDS], 2011). Details on the three main criteria parts: organisational infrastructure (section 3), digital object management (section 4) and infrastructure and security risk management (section 5), can be found in this thesis in chapter 4.2.

Bodies that apply TDR to a repository may also be conforming to the *Requirements for Bodies providing Audit and Certification*, a standard that was presented to the International Organization for Standardization in 2014. The requirements define the organisations that perform audits based on another ISO standard for certification of general types of management systems.

2.2.3.3 PREservation Metadata: Implementation Strategies (PREMIS) data dictionary

The PREMIS dictionary provides core metadata necessary for preservation of all kinds of digital objects and is currently in its second version. It is based on the OAIS reference model and can be viewed as a translation of the OAIS reference model into implementable semantic units (PREMIS Editorial Committee, 2012). Both standards share basic vocabulary even though PREMIS is more specific in some areas where the translation from the general OAIS framework into an implementable standard made it necessary. The PREMIS data dictionary defines five entities under which the metadata elements are grouped or linked: intellectual entities, objects, events, rights and agents. The goal of PREMIS is to maintain the information that is needed to identify, render and understand the data including:

- Who has owned the data, or where did they come from? (provenance).
- Information that determines the authenticity of the data.
- The preservation activities that have been taken so far.
- Technical considerations necessary to properly render the data.
- Who has the right to view, duplicate, etc. the data? (rights management).

The standard was published in 2005 and has been implemented by a variety of digital archives and libraries, to include the U.S. National Snow and Ice Data Center Repository (NSIDC) and its Data Model for Managing and Preserving Geospatial Electronic Records. Since the NSIDC considers it an internal project, there is not much information published about the PREMIS integration. Hoebelheinrich (2008) reports that PREMIS metadata can be used effectively for complex geospatial data under the condition that it is accompanied by domain-specific descriptions such as the FGDC metadata standard.

3 Approaches to geodata preservation

3.1 State of the art of geodata preservation

When searching international publications for projects in digital geodata preservation we came across very few scientific articles, some congress proceedings and mostly reports of current and past projects. Many actors that are involved with the preservation of geodata are not necessarily directly involved in a preservation project for geodata. This is the case, for example, for the Open Geospatial Consortium (OGC), which creates and promotes standards for geodata, which favour preservation, or EuroSDR, which produced a white paper on the topic, but is not involved as an organisation. Nevertheless, its members are involved in specific projects. We collected all documentation that was publicly available about the identified projects, to decide if they would further inform this research. The identified geodata preservation projects are presented in chapter 3.1.1. Chapter 3.1.2 follows the digital data life cycle introduced in chapter 2.2.1. For each step in the life cycle, actions taken by the identified legacy geodata projects that can assist long-term preservation are summarised.

3.1.1 Who is involved in the preservation of geodata?

We identified the following actors and their digital geodata preservation projects.

In the Americas:

- The National Oceanic and Atmospheric Administration (NOAA) of the United States of America in collaboration with the National Aeronautics and Space Administration (NASA).
- Several North American states and diverse institutions involved in a project called GeoMAPP.
- The City of Vancouver's preservation project called VanMap.

- The project 'National Geospatial Digital Archive' with a focus on collection undertaken initially by two and later four California universities.
- The state of Maine's GeoArchives.
- The Center for International Earth Science Information Network (CIESIN), with its focus as a legacy data clearinghouse.
- The Canadian Geospatial Data infrastructure GeoConnections, which is recommending the use of the OAIS standard, but does not describe an implementation itself.
- The State of California Archives, California Environmental Resources Evaluations Systems (CERES), which conducted the project eLegacy, selecting data from a geolibrary and using a testbed for preservation of the San Diego Supercomputer Center.
- The National Archives and Records Administration, which as receiving authority has developed a recommendation for scheduling potentially permanent geodata records.
- The Library of Congress as initiator of two reports and funder for several of the other projects.
- The National Snow and Ice Data Center, which published information about their selection and prioritisation process for collections.

In Europe:

- In the United Kingdom, the Ordnance Survey in collaboration with The National Archives (TNA) and several other partners – a service with a focus on access that provides use examples of legacy geodata on their website.
- In Holland, the Data Archiving and Networked Services (DANS) data archive as digital repository for research data; it also holds geodata.
- A Swiss collaboration between Swisstopo and the Federal Archives, which has developed appraisal criteria through a project called Ellipse.
- Several State Archives in Germany, which have analysed the permanent value of cadastre and mapping data. Projects are individually documented but the principal archivists of the German Länder published preservation guidelines through its Committee (ARK).

- EuroSDR, an organisation linking national cadastral and mapping institutions and research institutes. It developed archival recommendations for their members.

Cross continental actors with focus on geodata:

- The Open Geospatial Consortium (OGC) as a creator of standards and adviser to its members.
- The members of the European Space Agency and Canada that formed the Long-Term Data Preservation project (LTDP) (focus on the user).

For the purpose of this thesis, only initiatives specially focused on spatial data that answered one of the following criteria were retained:

- The project applies the OAIS standard to long-term digital preservation projects.
- The project focuses on the user.
- The project has developed appraisal or selection guidelines.

The resulting projects are presented in the following table. The brief presentation of the projects here does not include projects with a focus on collection and access such as repositories, catalogues and clearinghouses. The solution in the UK is such an example, with its focus on access. We could not identify application of OAIS in their system, however, the project implements PREMIS metadata.

Project	Partners	Year	Description	Output / focus
NOAA (no project name)	National Oceanic and Atmospheric Administration (NOAA) National Aeronautics and Space Administration (NASA)	2000-2002 for the joint project, ongoing for NOAA archives	The initial objective of their common curation effort was to preserve Earth Observing System (EOS) data. The decentralised archives at NOAA sites offer access for third parties and are characterised by their capability to adapt to fast increases in data holdings. http://www.class.ncdc.noaa.gov/	General data management and preservation instructions (Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA et al., 2007) and appraisal and selection guidelines (The National Oceanic and Atmospheric Administration [NOAA], 2008)
National Geospatial Digital Archive (NGDA)	Stanford University University of California Santa Barbara	2004-2009	The mission of the NGDA was to 'collect and archive major segments of at-risk digital geospatial data and images'.	The collected experiences focused on legal solutions for partnerships, collection development policies and format registry (Janée, Sweetkind-Singer, & Moore, 2009)
GeoMAPP	Archives, libraries and representatives of geosciences in Kentucky, Montana, North Carolina and Utah	2007-2011	Follow-up from the North Carolina Geospatial Data Archiving Project (National Research Council (U.S.). Committee on Research Priorities for the USGS Center of Excellence for Geospatial Information Science, 2007). Investigated several digital preservation issues, including business planning, data inventory and metadata, appraisal and access (Bethune, Lazorchak & Nagy, 2009).	The outcome of the project is published on the GeoMAPP website.[12] We would like to emphasise just three documents that were used in the context of the thesis: The Geoarchiving Self Assessment Tool, the presentation *Appraisal of Geospatial Data* and the documentation of the partner meeting on appraisal of geospatial data (Abrams et al., 2010)

[12] http://www.geomapp.net/

Project	Partners	Year	Description	Output / focus
Maine GeoArchives	The state of Maine Archives GeoLibrary Board	2004-2006	All governmental institutions in Maine are legally sanctioned to make their public spatial data accessible at the Maine GeoLibrary. The GeoArchives amends the existing infrastructure by adding a long-term perspective on the data access. The GeoArchives is a prototype for selected spatial data layers based on the commercial ESRI GIS.	Retention schedules for spatial data and procedural instructions for data transfer, appraisal and information flow (Henderson, 2006)
GeoConnections **No Project name**	Government of Canada GeoConnections	2001-ongoing	A Canadian program with the mandate and responsibility to lead the Canadian Geospatial Data Infrastructure through the use of standard-based technologies and operational policies for data sharing and integration.	*Recommendations on the Management and Preservation of Geospatial Data* (2004), *Geospatial Data Preservation Primer* (2013)
eLegacy	State of California Archives, California Environmental Resources Evaluations Systems (CERES)	2005-2010	Appraise and accession archival geospatial records held by the California Spatial Information Library (CaSIL), a division of the California Resources Agency and test a preservation testbed.	Project report with an explanation of the selection process (see document eLegacy in chapter 3.2.1)
NARA **No Project name**	National Archives and Records Administration		Geodata receiving agency.	*Tips for Scheduling Potentially Permanent Digital Geospatial Data* (see document NARA in chapter 3.2.1)
No project name	Library of Congress		Funded several of the previous projects through the National Digital Information Infrastructure and Preservation Program (NDIIPP) and commissioned two reports.	Study on the *Appraisal and Selection of Geospatial Data (2010)* *Issues in the appraisal and selection of geospatial data (2013)*

Project	Partners	Year	Description	Output / focus
NSIDC **No Project name**	National Snow and Ice Data Center (NSIDC)	Ongoing	Data clearinghouse and archive that implemented selection processes for data based on archival service levels. NSIDC involves users in the revisions of service levels and eventual retirement of data sets from the archive (National Snow & Ice Data Center, 2013).	Scientific articles on the selection and prioritisation process for collections to acquire (see documents NSIDC 1 and NSIDC 2 in chapter 3.2.1)
DANS **No Project name**	Data Archiving and Networked Services (DANS)	Ongoing	Research data clearinghouse and repository.	Selection and appraisal criteria for research data, including examples for geodata (see document DANS in chapter 3.2.1)
Project Ellipse	Swiss Federal Archives Swisstopo (Swiss Ordnance Survey)	2010-ongoing	Implementation of the Geoinformation Act that stipulates the transfer to and preservation of Swisstopo data at the Federal Archives.	Reports on the preliminary study (Bos et al., 2010) and presentation of the concept, describing the procedures including appraisal criteria (Projektteam Ellipse, 2013).
No Project name	Archivreferentenkonferenz (ARK) = (Committee of the principal archivists of the German Länder)	2006-ongoing	A sub-committee was engaged in analysing appraisal of data sets created by the state topological service, the choice of appropriate preservation formats and the transfer of selected data.	Publication of initial guidelines for accessioning geodata and making it available by public archives (see document ARK in chapter 3.2.1)
Long-Term Data Preservation project (LTDP)	European Space Agency and Canada (ESA)	2006-ongoing	The goal of the LTDP project is to harmonise the long-term preservation of earth observation space data between the member institutions. A common approach is needed to lower the cost of preservation and ensure accessibility and usability of the collection. LTDP promotes its guidelines as strong recommendation to its members	To our knowledge it is the only preservation project having published a user study to determine the user needs for legacy space geodata (Molch, Leone, Albani, & Mikusch, 2012)

Table 1: List of geospatial data preservation efforts and involved partners with outputs related to this thesis

All projects are partnerships between data producers and storage institutions and insist on the utility of collaboration because of decentralised production, data ownership and management and technological expertise. The geodata producers contribute with understanding about file formats, immediate use of data and specialised topic knowledge. Heritage institutions comprehend long-term risks, preservation processes and description standards, and they know about the diversity of possible future uses. Because many producers do not focus on long-term access, they ignore use cases of legacy geodata.

3.1.2 How to preserve geodata? Challenges and different approaches to curation and preservation

The field of geodata preservation is not so far along as to offer best practices for specific fields like cartography. Lack of experiences in long-term archiving by mapping institutions is perceptible in the literature. Many of the project reports analysed come from institutions that deal with satellite imagery, which is not the scope of this investigation. Nevertheless, the technological challenges that satellite imagery present are similar to aerial photography as both are forms of capturing electromagnetic radiation at different altitudes (Longley et al., 2015). Projects focusing on preservation of satellite remote sensing have therefore been taken into account in the extraction of the preservation challenges. In this chapter, the special challenges to geodata preservation that were mentioned in the analysed project reports and scientific literature are synthesised. The sub-chapters present proposed solutions, contexts and actions that can be taken during the whole geodata life cycle and that favour long-term preservation. Although it is not named in the data life cycle model we name data management first, because it is supposed to accompany data from its creation to its preservation. Thereafter we follow the documentary chain of content creation – description – appraisal and selection– ingest – storage –access and transformation, because measures for improving geodata preservation can be introduced at any of these steps.

3.1.2.1 Data management

Good data management eases digital preservation. In the geosciences, it is more common to speak about data policy or data management policy (Steinhart, 2006). A good data policy, from an archival perspective, includes end-to-end data management. This would mean that during planning for data creation, budgeting decisions would take into account data preservation actions that are necessary after project life.

The International Council of Scientific Unions (ICSU) developed data management plans in 1957-1958 for each scientific discipline that was part of the International Geophysical Year (IGY) project. We can read about early data management plans on the World Data Centers website:

> *Multiple Centers were established to guard against catastrophic loss of data, and for the convenience of data providers and users. The IGY planners were remarkably prescient: the 1955 recommendation mentioned that Data Centers should be prepared to handle data in machine-readable form, which at that time meant punched cards and punched tape.*[13]

The open data initiative has taken over this idea of data sharing. It calls for making data accessible through electronic infrastructures such as the internet.

One of the most influential U.S. institutions in geosciences data archiving is the National Oceanic and Atmospheric Administration (NOAA). The NOAA report (Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA et al., 2007) advises producers to manage environmental data end-to-end. This means beginning with the creation of the data through the storage and preservation of data sets. Nevertheless, to date, it is exceptional for a spatial data producer to have policies that include digital data preservation. Harris (Harris, 2001) distinguishes two different basic situations equally problematic for the long-term survival of data: data collected during a project-based activity (such as research in a university or satellite mission) and data sets produced during recurring capture by institutions with regular funding and a legal mission. Harris points out that remote sensing data collected by satellites usually belong to the first group and are preserved during the mission of a satellite and sometime more. Project-based data collection rarely includes funding for data preservation, whereas data coming from the second group are ingested in GIS programs for access and most likely overwritten. Only over the last 10–15 years have some government archives started to take over geodata for preservation (see chapter 3.1.1).

[13] http://web.archive.org/web/20090309004839/http://www.ngdc.noaa.gov/wdc/about.shtml

Challenges: Data management plans and policies are first and foremost developed to support efficient compliance with the primary purpose. It is challenging for archives that have no other contact with geodata producers to communicate their needs and eventually make producers change policies.

Advantages and opportunities: The leadership of influential institutions and renowned international endeavours bring recommended best practice for good end-to-end management closer to smaller producers. General enterprise quality management increases trust in the data and eventually to data completeness and accuracy. The Science Data Infrastructure for Preservation-Earth Science (SCIDIP-ES) report *Parameters for Long Term Preservation and Data Sustainability Models* (Caruso, Briguglio, Matthews, & Polsinelli, 2013) identifies ‘quality management’ as one of the supporting activities for long-term preservation.

3.1.2.2 Data creation

Data creation is strongly decentralised in the geosciences. Public and private institutions produce data on all levels, from local government to international institutions. Additionally, many institutions licence data from others to do data enrichment. The final geodata product might contain data from different sources with various rights. Data from international projects such as ESA or Landsat imagery are distributed and stored in various countries (Beruti et al., 2010). Maps are produced on different scales, depending on the organisation’s authority in a certain region. The continuation of the same map scale beyond the border might be created and owned by another producer, adding semantic problems such as explained in the discussion about tiles (chapter 2.1.2.2). Obtaining data is quite expensive because of necessary measurement instruments, human resources and transport for sensing.

Challenges: Preservation plans and collection development have to include related data created or stored in different sites to be able to offer a continuous and coherent data set. Additionally, rights from a variety of producers might complicate data transfer and access. The fact that most geodata are inherently digital means that they might lose significant properties when migrated to other formats, as is common in digital preservation.

Advantages or opportunities: The cost of data creation can be a criterion in favour of preservation.

3.1.2.3 Data description

Data description is done on different levels and on different parts of data to serve specific purposes. Information about digital objects is also called metadata:

- Description at the bit level that is collected in the representation information is important to render data in the future.
- Description of the intellectual content aims to improve discoverability and allows evaluation of suitability for reuse purposes.
- Description of preservation processes such as in the PREMIS standard serves data management, such as for preservation actions and access.
- Description of the preservation unit is done to check completeness, corruption and/or conformity with given requirements.

We would like to focus on descriptive information for the intellectual content. In the OAIS model, the metadata that describes the intellectual content is part of the collection description. Even though it could be a free-text description, the use of machine-readable metadata standards has significant advantages for the harvesting and discoverability of data sets (Powell, Heaney, & Dempsey, 2000). We need to know if a line represents a fence or a street, if colour depth represents the concentration of a measured value or geodesic height and which is the unit of measurement of numeric data, the projection or the technical specification of the data capture, etc. If underlying data does not contain its own metadata or if only the representation is available, additional metadata is needed.

The encountered metadata solutions listed below state that the description of the data set is usually made in XML format,[14] as a separate metadata file or integrated in the geodata file format. As awareness of preservation and access rises, guidelines on data description for producing institutions are emerging.[15] These guidelines recommend using standards for content description. The most popular description standard intended for data access is

[14] From the Federal Geographic Data Committee metadata definition: http://www.fgdc.gov/metadata/index_html

[15] For example, *Managing Historical Geospatial Data Records: A Guide for Federal Agencies* (http://www.fgdc.gov/library/factsheets/documents/histdata.pdf) or the *GeoMAPP best practices*.

Dublin Core. Because it is not specifically for geodata, we omit it in the following list of metadata standard descriptions.[16]

3.1.2.3.1 Metadata standards for geodata

The most used metadata options for content description especially designed for geographic data are presented as follows.

ISO 19115:

The International Organization for Standardization maintains the ISO 191xx family on geographic information, of which ISO 19115 is the spatial metadata standard. With its over 400 metadata elements and 20 defined core elements, its goal is to enable the description of a spatial data set in a way that makes it possible for users to evaluate the data relevance. The standard is divided in two parts. Part 1 was originally published in 2003 and revised in 2014 as *ISO 19115-1: Geographic information – Metadata – Part 1 Fundamentals*. Part 2 was published in 2009 and adds extensions for imagery and gridded data. The same family of standards provide XML schema implementations for the two parts that specify metadata record format.

The United States and Canada have developed an application profile called the North American Profile for ISO 19115, which is currently under revision. ISO 19115 and the North American Profile replace the former Federal Geographic Data Committee's (FGDC) standards and profiles. The FGDC now recommends that U.S. public and private spatial data creators use the NAP.[17]

FGDC Content Standard for Digital Geospatial Metadata (CSDGM)

The CSDGM is also known as the FGDC metadata. Although it has been superseded, many repositories still use it, because it was made mandatory in 1994 for publicly funded geospatial projects in the United States to create FGDC-compliant descriptions. Due to its implementation in American GIS technology, which controls most parts of the market, the standard reached worldwide use. It was intended to aid the discovery and identification of

[16] For further information on Dublin Core please refer to the Dublin Core Metadata Initiative website: Dublin Core Metadata Initiative (http://dublincore.org/)

[17] Federal Geographic Data Committee. *Geospatial Metadata Standards and Guideline.* https://www.fgdc.gov/metadata/geospatial-metadata-standards/

data sets. The structure of the former CSDGM maps to the newer ISO 19115, which eases data migration and interoperability.

Global Change Master Directory Interchange Format (DIF)

The Global Change Master Directory (GCMD) acts as a catalogue of spatial data sets. The descriptions are held in the Directory Interchange Format (DIF), a special description format created in 1988 by the geosciences community in the United States. The DIF is maintained by NASA[18] and adopted elements of the CSDGM in 1994. After ISO 19115 became effective, required elements and appropriate modifications were incorporated into the DIF to achieve full ISO compatibility.[19] A description that follows the DIF rules version 10 has thirteen mandatory fields among which are 'temporal coverage' and 'spatial coverage'. Several fields expand upon and clarify the information of others. Some of the fields require the use of controlled vocabulary, such as the 'science keywords' field.

Infrastructure for Spatial Information in the European Community (INSPIRE)

The INSPIRE metadata standard is the European norm for geospatial data description, valid since 2008. It has been specified in the *Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata* and is part of an effort to create a data sharing infrastructure (European Union, 2007). INSPIRE defines 31 metadata elements divided in 10 sections for spatial data sets, spatial data series and spatial services:

- Identification
- Classification of spatial data and services
- Keyword
- Geographic location
- Temporal reference
- Quality and validity
- Conformity
- Constraint related to access and use
- Organisations responsible for the establishment, management, maintenance and distribution of spatial data sets and services
- Metadata on metadata

[18] U.S. National Aeronautics and Space Administration

[19] http://gcmd.nasa.gov/User/difguide/whatisadif.html

The technical guidelines of the implementations rules map each INSPIRE metadata element with ISO 19115 and ISO 19119 core elements. INSPIRE is more demanding in some elements but it is missing other elements compulsory by the ISO norms. Therefore, an INSPIRE description does not automatically fully conform with ISO 19115, nor the other way around, although ISO 19115 was the basis for the development of INSPIRE (Litwin & Rossa, 2011).

Data Model for Managing and Preserving Geospatial Electronic Records (GER)

The Center for International Earth Science Information Network (CIESIN) has developed the GER data model to better manage electronic spatial data with the scope to include the entire life cycle. Version 1 was published in 2005, and is free to use for the science community. The GER data model is optimised to support selection, accession, preservation and disposition of the data but is not intended primarily to serve discovery and access (Center for International Earth Science Information Network [CIESIN], 2005). Nevertheless, it offers a crosswalk to Dublin Core. The first part of the GER model is the entity-relationship diagram, which identifies five groups of metadata entries:

- Provenance and attributes
- Organization
- Distribution
- Administration
- Physical properties

Each of these groups is composed of several tables with metadata entries. The entity-relationship diagram shows the relationship the elements have with each other. The second part of the GER model is the data dictionary, which provides definitions for each metadata element of a table. Even though it is not intended to be PREMIS-compliant, the GER model presents a crosswalk to the PREMIS metadata and several other preservation and management metadata standards. It is specifically strong in taking into account different kinds of geographically related information, such as addresses for example. Nevertheless, there is no documented implementation of the GER data model and since the end of the initial funding it has not been developed further.

GeoMAPP Archival Metadata Elements for the Preservation of Geospatial Data Sets

Specific metadata elements for geospatial data based on the OAIS model have been proposed by the GeoMAPP project (Geospatial Multistate Archive and Preservation

Partnership [GeoMAPP], 2011a). The elements have been taken from three different sources: the GER data model,[20] the CEDARS project preservation metadata elements[21] and Hoebelheinrich's study (Hoebelheinrich & Banning, 2008). The GeoMAPP project maps those metadata elements to the basic OAIS information elements and therefore exemplifies the OAIS reference models metadata for geospatial content. This metadata set unites elements intended to serve preservation with those for access in one standard.

3.1.2.3.2 Value standards for metadata elements

Value standards regulate what can be written into a metadata field. In some cases, a description standard element refers to other standards when it is important that a certain value be controlled. In other cases, the use of value standards is optional. Because their use helps understanding and identification of the data but touches preservation only on the fringes, we describe two of them very briefly.

Unique Identifier

A unique identifier is a character chain that is different for each object of a set, such as the ISBN number for books. For internal uses, unique identifiers valid within the producer or collector site are sufficient. However, internal identifiers with the same value that come from different producers might cause problems when data are exchanged between them. Universal validity would assist data interoperability between institutions. Identifiers obtained from internationally recognised institutions should be included in the metadata of a data set. DataCite is such an institution; it assigns persistent identifiers to data sets.[22] The PANGAEA[23] earth and environmental science publisher and the Dutch DANS[24] use digital object identifier's (DOIs) for each data set. DOIs are used successfully and very popular for electronic journal articles. If the data are on the web, Uniform Resource Identifiers (URIs) can be used. NOAA states that their archived data will soon be provided with unique identifiers (De La Beaujardière, 2016). The mentioned unique identifiers are applied on the level of data sets. We can expect that with further integration of data, the need for finer granulated identifiers will emerge.

[20] Data Model for Managing and Preserving Geospatial Electronic Records (GER)

[21] http://www.ukoln.ac.uk/metadata/cedars/papers/aiw02/

[22] http://datacite.org/whatdowedo

[23] http://www.pangaea.de/about/

[24] https://dans.knaw.nl/en/deposit/information-about-depositing-data/persistent-identifiers

Ontologies and Dictionaries

There can be dictionaries, also called controlled vocabulary, for various type of information: place names, thematic subjects, etc. The reference to the real-world location of a digital object is expressed in coordinates of longitude and latitude, but the metadata can also contain place names. They might be extracted from a controlled list called a gazetteer. A gazetteer is a geographical dictionary that serves as reference for place names. Another example of a controlled vocabulary is the GEneral Multilingual Environmental Thesaurus (GEMET). GEMET is an effort of some European countries to provide consistent indexing in a multilingual context.[25] Instead of being tied to a dictionary, a metadata field can be tied to an ontology with complex structured concepts. More on ontologies can be found in chapter 3.1.2.6.1.

Challenges:

There are many metadata standards available and in use. Variety is an obstacle to integration or identification of data sets. Nevertheless crosswalk efforts between metadata standards allow for legacy data being discovered through spatial data infrastructures designed for current geodata (Capdevila Subirana, Sánchez Maganto, Camacho Arranz, & Arístegui Cortijo, 2012). Quality metadata that is useable beyond institutional borders or research domains is costly to produce. Schemes where all costs fell on the producer and typically any benefit accrued to the users have had little success (Longley et al., 2015).

Advantages and opportunities:

Data description is very important to preservation because it makes data sets interpretable for future use (S. P. Morris et al., 2009). Uniform metadata at the content level assists data set inventories and makes them discoverable for possible selection. Standardisation is underway and increases the probability of development and success of additional tools, such as metadata validation.

3.1.2.4 Data assessment and selection

Before it will preserve geospatial data, an institution has to have sufficient motivation. An institution interested in maintaining data for its historical, economic or scientific value can

[25] http://www.eionet.europa.eu/gemet/about?langcode=en

proceed to active assessment of the value. Assessment might involve elimination of data sets judged valueless. Selection might be additionally motivated by limited resources (space, funding, infrastructure, etc.). Selection becomes pressing especially for big data sets such as remote sensed data (Janée, 2008). Satellites often create huge amounts of data, so that storage space becomes a problem. The amount of remote sensed data by ESA satellite alone reaches two terabytes a day, above the capacity of most data archives. (Harris, 2001).

Chapter 3.2.1 will give an overview of many geodata assessment guidelines and reports currently available.

Challenges: Even though NOAA has created an influential appraisal policy (The National Oceanic and Atmospheric Administration [NOAA], 2008) that formed the basis of discussion at the Library of Congress geospatial appraisal meeting in 2010, we are far from having a universal appraisal guideline for geodata.

Advantages and opportunities: Space and resource problems make producers sensitive to digital preservation solutions. Assessment of data at different points in the data life cycle can inform appraisal for preservation.

3.1.2.5 Data storage

One challenge of digital data preservation is making it clear to decision-makers that simple storage and backup is not enough for long-term preservation. But the principles of backup remain: making multiple copies and storing them in geographically diverse sites lowers the risk of data loss. Even storage can be a problem when large amounts of data are in the mix. Geoscience institutions produce high numbers of digital photographs with very high resolution, which occupy a lot of storage space. Big data has been an issue since the beginning of computing; we can see it mentioned in the manual *Automatic Data-Processing Systems: Principles and Procedures* from 1960 (Gregory & Van Horn, 1960). But in the last few years it has become the main subject of some publications. When storage media is costly, new data is written on older backups. NASA supposes that this was the way it lost Apollo 11 tapes of the first human on the moon (NASA, n.d.).

Because of data size, data producers share sensing tasks and divide areas or subjects between members. The same reasons will force archiving institutions to share preservation responsibilities and encourage them to work in networks. This way, they will be able to store more data and to assure redundancy on geographically different sites. Data are at risk whenever they are stored in only one geographical location, even if there are several copies of the data. Networks, as the NGDA project shows us, can allow for remote storage by partner institutions.

Spatial data are stored in different stages. A basic distinction is made between raw and processed data (The National Oceanic and Atmospheric Administration [NOAA], 2008). Processing can be minimal or data can be elaborated to complex products. Processes might include quality control, calibration, data display through GIS, generalisation, etc. The Ellipse preservation project started with the assumption that the processing stages of geodata could be divided into the following status: production data, official data and commercial services. Production data include raw data, measurement data and observations. Official geodata comprise adjusted measurement data, production data or working data. Commercial services or products include such things as maps, evaluations in the form of statistics and diagrams. Application of this classification in practice shows that it cannot be applied to all geodata, as official geodata might exist only as measurement and not all data make it to the stage of a commercial service. The distinction of data in their different process stages will play a role for appraisal and selection, because they can be seen as duplicated data. For accomplishing their tasks users might replace data of one process stage by another stage. Instead of using the term process stage DANS refers to stages in the digital life cycle and also refers to raw and processed data. Some researchers publish the data their research is founded on. As compared to published governmental map data, the publication of research data does not correspond to an additional process stage, because it is not further transformed for publication. It is considered by DANS as an indicator of data significance or quality.

3.1.2.5.1 Versioning for the archive

Versioning is defined as the management of recurring capture of information. We must distinguish between two slightly different understandings of versioning in geoscience and information management. The geographic point of view and technical specifications are explained in chapter 2.1.3. In information management, when talking about archiving databases, the version is the ‘snapshot’ of the state of the data at a certain time that makes it into the archive. Archives and libraries might create archival versions of databases, based on the existing versions at the producer. Erwin calls the decision about the frequency of archiving older or changed data versioning (Erwin et al., 2009). When no versioning is practiced at the producer, but the archive has the authority to capture snapshots, archival versions are the only evidence of historic states. When the snapshot approach is practiced by the producer, the acquisition interval of the archive can be equal to the snapshot frequency of the producer or more spaced in time. Therefore, the versions held by the data producer are not necessarily the same as those finally archived. It is shown that object- and event-oriented approaches to versioning create large databases that might exceed the capacity of an archive, if regular updates are made (Gantner et al., 2013). Therefore, object- and event-oriented databases might be transformed to a simpler database structure for archiving.

The frequency of archiving is documented by the retention schedule. Erwin (Erwin et al., 2009) and Dingwall (Dingwall et al., 2005) agree that decisions about the acquisition interval must be taken for each data type individually. Dingwall uses the map layers of VanMap to explain: The layers need to be evaluated within the context of the business processes they support. Although the general trend is to archive layers each time they are updated, there might be changes insignificant to the decision process the layer supports, such that they would be archived at a lower frequency.

Thematic geodata present an additional challenge in versioning: the thematic layers depend on reference data (underlying map) that can be actualised with different frequencies. The Swiss Federal Archives has analysed three scenarios for the specific problem of thematic spatial data (Bos et al., 2010):

- All changes in the reference data are captured to assure that the thematic data always have their corresponding reference. This might result in capturing reference data that are not needed.
- Every time a thematic data set is captured, the corresponding reference data are archived as well. This might result in redundancy of reference data.
- Thematic and reference data are captured at a fixed frequency. This might result in thematic data that miss the correct reference data.

Because all of the three variations have their disadvantages, Bos et al. recommend using a combination.

The frequency of archiving is important in regard to legal obligations where an archived record must reproduce an exact view of the data, as they were seen at a certain time. If data have such a legal function, each iteration may need to be preserved. This is the case, for example, to prove the date a change in a city's infrastructure was made (Sweetkind et al., 2006).

3.1.2.6 Data access and reuse

Before data can be reused they must be discovered by potential users. Data set inventories were created with the idea of making data discoverable for sharing or for their selection and acquisition. Some national inventories are mandatory, like the official spatial data registries in Spain,[26] others are voluntary or even created with a commercial purpose. Some inventories specialise in spatial data, like the Coastal and Marine Spatial Planning (CMSP) Data Registry.[27] Others are forged for all science data, like the UK Data Archive.[28] The inventories describe data at the data set level and assure that data sets can be localised and potentially reused. Registries, data infrastructures, inventories and clearinghouses use metadata standards for description. Some are discussed in chapter 3.1.2.3.1.

Once data are discovered and judged relevant, a user will want to access and use them.

In science, the term data exchange is more common than 'reuse'. Information exchange has always been important to verify statements or to test new methods on previous knowledge. Exchange is practiced between universities, laboratories, government organisations and professional associations. Many topics influence data exchange: infrastructure, interoperability, rights management and data quality. The determination to share data can fail due to missing logistical and technological infrastructure. In the geoscience field, the International Geographic Year project built links between research institutions and tightened cooperation, favouring this way of data sharing in the geosciences after the Cold War. Preservation capitalises on data exchange because it forces the partners to foster data description and to use standards (e.g., for file format and metadata), and it creates networks between institutions on which preservation efforts often depend. Research itself benefits from data sharing, because it allows interdisciplinary studies (Moran et al., 2009).

[26] http://www.01.ign.es/ign/layoutIn/actividadesRcc.do

[27] http://egisws02.nos.noaa.gov/cmspgisdataregistry/

[28] http://www.data-archive.ac.uk/create-manage/strategies-for-centres/data-inventory

Advantages and opportunities: It is common sense that frequently used or requested digital data are less likely to disappear. When stakeholder interest is manifested, data and rendering environments might be kept available.

3.1.2.6.1 Technical infrastructures

On the European level, governmental data sharing has been reinforced by *Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information* (European Union, 2003), which obliges public institutions to publish any kind of data created with public funds. In addition to this legal incentive, the nations of the European Union are currently implementing the INSPIRE infrastructure for spatial data (European Union, 2007). Apart from the metadata standard, already discussed in chapter 3.1.2.3.1, INSPIRE's goal is to make geodata available through a common infrastructure. The INSPIRE directive obliges the participating countries to share public sector environmental information and aims to ease access for the general public. Implementation is planned to terminate in 2019. Similar projects have been realised in the United States, where in 1994 a plan for coordinated access and acquisition of geospatial data led to the creation of the National Spatial Data Infrastructure (NSDI). Executive Order 12906 (United States of America, 1994), which is the legal base for the infrastructure, assigns coordination of the effort to the Federal Geographic Data Committee and contains specifications about geospatial standards.

3.1.2.6.2 Interoperability

We often hear about data interoperability, which is related to data sharing. Interoperability is understood as the ability to develop conventions to make data exchange and integration possible (Nogueras-Iso, Zarazaga-Soria, Lacasta, Béjar, & Muro-Medrano, 2004). There is the technological aspect to interoperability dealing with file formats and there is the logical aspect, dealing with semantic interoperability. Technical interoperability is the capacity of data to be interpreted and reused in different technological environments without it having to be translated or migrated into a new format. Technical interoperability allows us to open, for example, Esri shapefiles in geographic information systems not designed by Esri. To be interoperable, two data sets or file formats must share technological properties responding to appropriate standards. 'Semantic interoperability provides systems with the ability to access, consistently and coherently, to similar (though autonomously defined and managed) classes of digital data, objects and services distributed across heterogeneous repositories' (Nogueras-Iso et al., 2004, p. 612). Disciplines have developed specific metadata standards to suit their data. As metadata descriptions relate and overlap semantically, crosswalks between the standards are developed. According to the *Dublin Core Metadata Glossary*,[29] a crosswalk is 'a table that maps the relationships and equivalencies between two or more metadata schemes. Crosswalks or metadata mapping support the ability of search engines to search effectively across heterogeneous databases'. Current trends introduce ontologies at the data model level. The structure of such data models dissuade from the use of relational databases for those data (Gantner et al., 2013). Profoundly different is the ontology-based metadata approach that is straightforward to implement and easy to understand by GIS users. Ontology-based metadata can bridge the discrepancies between two metadata standards and accounts for automatic semantic integration across discipline boundaries (Schuurman & Leszczynski, 2006).

Even when interoperability is provided, data exchange can still be difficult because of data size. Where infrastructure is missing or lacking, capacity exchange can take the form of shipping hard drives between institutions, as experimented with in North Carolina and the National Geospatial Digital Archive.

Challenges: Existing data infrastructures only cover a small part of the global production. For example, the INSPIRE initiative covers only European environmental data. Diverse producers are not concerned to comply and implement INSPIRE. Semantic interoperability is in opposition to cultural differences.

Advantages and opportunities: Transfer from the producer to the archive or between producer and user benefits from existing exchange infrastructure (S. P. Morris et al., 2010). Infrastructures help archives discover relevant data sets for archiving.

3.1.2.6.3 Rights management

Information about the legal rights to geospatial data is very important. Future viewing and reuse depend on the correct handling of the rights information, as does managing of data in the archive. In preservation systems, rights information can be stored in metadata. In

[29] http://dublincore.org/documents/usageguide/glossary.shtml#C (last updated on 23 April 2004).

commercial data management systems, a whole rights management component can be implemented. Data custodians have to be aware that topic geodata layers often depend on reference layers licensed from other owners (Mcgarva et al., 2009). The rights of both, the base layer and the topic layer, must be documented. Data producers that want to distribute their data without losing control of the rights can develop applications using the Geospatial Digital Rights Management Reference Model (GeoDRM RM) elaborated by the Open Geospatial Consortium. The goal of this reference model is to enable applications to work with data under public licences, such as the Common Public License or the General Public License, and to manage payments for commercially licensed data sets. In 2014 the Digital Rights Management Reference Model became international standard ISO 19153:2014.

Challenges: Archives are not necessarily informed about the various licences of a combined data set, but need this information to provide lawful access to data. Data sets with various licences bedevil access and can make processes unsustainable for the archive. When related data cannot be obtained, certain dependent data risk losing interpretability.

Advantages and opportunities: The correct implementation of the INSPIRE metadata should inform the archive about provenance of all parts of the data set. Knowledge of previous owners can assist in obtaining necessary licences.

3.1.2.7 Data transformation

The data life cycle suggests data transformation for reuse. Transformation can occur for various reasons. Ideally, data are offered in several file formats for the user to choose from. To reach this level of service it might be necessary for the archive to transform data from the original or archival format to a currently used and accepted file format. Reuse rarely has the same purpose as initial use and might therefore also involve data transformation in the sense of partitioning. A data set might cover a larger area than is needed or contain more layers or features than desired by the re-user.

Challenges: Resubmission of transformed data might increase redundancy when the new or original data cannot be usefully separate from the already archived part.

3.2 The role of appraisal and selection in preservation

There is no clear distinction between the use of the terms appraisal and selection, although 'appraisal' is more current in archival science and 'selection' is more often related to library practice (S. Morris, 2013). Many times both terms appear in the same phrase, such as in the reports and papers of Abrams et al. (2010), Mcgarva et al. (2009), S. P. Morris (2010) and Tjalsma and Rombouts (2011). In the view of an archive, appraisal is the assessing of material for its long-term value. It involves selection of material for archiving out of a too-vast amount of information. Appraisal can be understood as the assessment, while selection would then be the separation of what is going to be archived from the material to be left out (Tjalsma & Rombouts, 2011). The goal of appraisal and selection is to keep as much as necessary and as little as possible. The necessity to maintain the understandability of data and the context of their creation, and to uphold their potential to be combined, lends support for including data in the collection while the necessity to keep the effort and resources for archiving them on a sustainable level lends support for excluding data from the collection.

Appraisal is about assigning value. The concept of value may be understood as the relationship between the user's perceived benefit and the effort for receiving these benefits (Caruso et al., 2013). In the early phase, data have administrative, informational and legal value. With time, some archival documents lose their administrative and legal value and gain social, historic and cultural value. The archive that keeps data hopes to create, not monetary, but other forms of value in return. The potential of geodata to be combined with data from other topics provide them with scientific value that they do not lose throughout their life cycle. Cultural, social and historic value is not assessed by how costly it is to produce or reproduce the data, although the cost of production can be a risk factor, if data are not preserved and therefore must be reproduced. Therefore, production cost considerations sometimes flow into appraisal recommendations. Because basic geodata in general have location and time references, they cannot be reproduced subsequently. Production cost can play a role only when assessing geodata derived through automatic processes from preserved basic data.

Appraisal of material that is intended to be included in an archive is important for several reasons. As said before, data are often considered to be intermediate data. Intermediate digital data usually lack understandability by people who are not involved in their production. One of the goals of the OAIS is to guarantee understandability and usability by the designated community. If too much effort has to be invested to achieve understandability, it is not sustainable for the archive to keep the data. Furthermore, digital data are very susceptible to being copied. Copies that are not specifically purposed to long-term preservation should be eliminated, because they occupy valuable storage space and devour resources when preservation actions have to be applied on the collection.

Appraisal can be situated at several levels of the archive's acquisition process (Tjalsma & Rombouts, 2011). The earliest point at which we can talk about appraisal is when the producer has an agreement with the archive to evaluate the long-term value of data at the point of creation of a new series. This is only the case when creator and archive work very closely together. When archives are involved late in the data life cycle they might have to appraise data stored on obsolete or even physically damaged media. The point at which appraisal is done dictates whether criteria on physical conditions of storage material can be part of the assessment. Early appraisal at the planning stage of a new data series can only take into account content- and context-based criteria, which are usually more open to interpretation. Considerations about technical feasibility of preservation are then not part of appraisal but of planning.

An archive might want to evaluate if a series will fit into its collection in a similar manner as would a library. This type of appraisal can be triggered by a collection that is offered to an archive that is not legally obliged to take that collection. This happens, for example, at the U.S. National Snow and Ice Data Center which receives many requests for archiving (Duerr, Weaver, & Kaminski, 2010). Appraisal can also be initiated by an archive to enhance its collection with relevant documents. This procedure is more often called selection. Appraisal by legal incentive is done mostly for series as a whole, and is done only once, as long as the production process of that series stays the same. One appraisal decision can then be valid for many years and several data transfers.

Sometimes the term appraisal is used only for the evaluation of the intellectual content. In these cases other processes involve assessment of technical, legal and other constraints because they can be critical to correct preservation and access. Ideally, either the necessary effort for reaching quality data is spent at ingest or data have gone through previous formal checks. Individual concepts in the OAIS reference model are kept flexible so that appraisal can include parts of ingest procedures such as quality checks or estimation of storage space growth of the collection. Appraisal can result in rejection by the archive if required data quality is not reached or if estimated storage space growth is unmanageable for the archive.

3.2.1 Appraisal guidelines in comparison

Because value is by definition subjective, archives establish criteria to guide the decisions of the individual archivists and to develop a homogeneous collection. Nevertheless, most criteria leave room for interpretation and therefore appraisal can lead to heterogeneous results. Decisions should be as homogeneous as possible; especially when data sets that depend on each other or that are related in some way are appraised separately. Heterogeneous decisions on a horizontal level between institutions can lead to discontinued coverage, on a vertical level at the different scales of government such decisions can lead to discontinued resolution levels, and between related topics they can lead to loss of reference for the thematic data.

Following is a list of the guidelines we compared, preceded by an abbreviation chosen for further reference to the related document:

EROS:

The Scientific records collection appraisal question set by the North American Earth Resources Observation and Science (EROS) project (Last accessed 11 January 2016). The EROS Center is a U.S. Geological Survey (USGS) facility with a national and international mission to enhance understanding of a changing earth. It has legislative charters to preserve and make land remote sensing records accessible. The proposed appraisal questions are designed for a repository dealing with large amounts of geodata offered by third parties. Because it was developed by the standards-giving body USGS, the EROS question set influenced different other guidelines.

eLegacy:

Guidelines from *Appraising Geospatial Records: Strategies and Guidelines*, a report from the project, California's Geospatial Records: Archival Appraisal, Accessioning, & Preservation. Sustainable Archives and Leveraging Technologies, (California State Archives, The University of North Carolina at Chapel Hill, & California natural resources agency, 2010). The appraisal guidelines of this project are focused on selection criteria for the project itself, which focused on automating the accessioning process for data sets into a given preservation testbed.

LoC:

Appraisal and Selection of Geospatial Data by Steven P. Morris. Library of Congress (2010).

NDSA-LoC:

Issues in the Appraisal and Selection of Geospatial Data: an NDSA Report by Steve Morris. National Digital Stewardship Alliance - Geospatial Content Team (2013). Available online: http://lcweb2.loc.gov/master/gdc/lcpubs/2013655112.pdf

DCC:

Curating Geospatial Data by Guy McGarva. DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Center (2006). Available online: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

NSIDC-1:

Maintaining Data Records: Practical Decisions Required for Data Set Prioritization, Preservation, and Access by R.L. Weaver, W.M. Meier and R.M. Duerr. Proceedings of the 2008 IEEE International Geoscience and Remote Sensing Symposium. This article develops the product maturity model from NOAA.

NSIDC-2:

Data Acceptance Procedures and Levels of Service at the National Snow and Ice Data Center by R. E. Duerr, R. L. Weaver, M. Kaminski (2010). This document is not about appraisal, primarily. It addresses prioritisation decisions about levels of service and procedures when data are offered to the archive. It refers to NSIDC-1 for criteria.

Ellipse:

Concept for the Archiving of Official Geodata under Federal Legislation. Section D – Appendix 15 and 16 elaborated by Project Ellipse (2013). Available online:

https://www.bar.admin.ch/dam/bar/en/dokumente/konzepte_und_weisungen/konzeptbericht_ellipse.pdf.download.pdf/concept_report_projectellipse.pdf

Appraisal is a shared responsibility between the producer (Swisstopo) and the archive (Swiss Federal Archives). Some criteria are those of the archive that receives the data and some are of the creator, who knows best the production and its technical difficulties, intellectual content and relation.

NARA:

Tips for Scheduling Potentially Permanent Digital Geospatial Data Records developed by NARA. Retrieved June 14, 2017 from https://www.archives.gov/files/records-mgmt/publications/geospatial-tips.pdf

LTDP:

Long Term Preservation of Earth Observation Space Data: European LTDP Common Guidelines. Annex 3 – Data purging alert and data appraisal procedures. This guideline recommends using the EROS questions or the guidelines of the Committee on Earth Observation Satellites Working Group on Information Systems and Services that currently refers to general NARA Electronic records appraisal questions and the EROS tool. These references are not present any more in the newer version of this document published in 2015 (Ground Segment Coordination Body, 2015) which is why the earlier version of 2012 is used.

ARK:

Handreichung zur Archivierung elektronisch vorliegender Geodaten by Beate Dofey et al. of the Archivreferentenkonferenz (Committee of heads of archives in Germany) (2009).

GeoMAPP:

Appraisal of Geospatial Data. Presentation published by the GeoMAPP project (2011). This document is based on NARA recommendations and the EROS appraisal tool.

Kentucky:

The state agency records retention schedule of the Kentucky Department for Libraries and Archives (n.d.).

Maine:

Creating the GeoArchives: Maine Archives of Geographic Information by Jim Henderson. Report presented at the joint conference of NAGARA, SAA and CoSA (2006).

GeoConnections:

Management and Preservation of Geospatial Data by D.L. Brown, G. Welch and C. Cullingworth. Published by the Ad-Hoc Committee on Archiving and Preserving Geospatial Data in Canada (2004).

NOAA:

NOAA Procedure for Scientific Records Appraisal and Archive Approval: Guide for Data Managers by the National Oceanic and Atmospheric Administration (2008).

DANS:

Selection of Research Data: Guidelines for Appraising and Selecting Research Data by Tjalsma and Rombouts. Data Archiving and Networked Services (2011).

NGDA-MIL:

Collection Development Policy for the National Geospatial Digital Archive node, Map and Imagery Laboratory. Map and Imagery Laboratory (2007).

An additional appraisal guide, *Guidance on the Selection and Appraisal of Geospatial Content of Enduring Value* elaborated by the Users/Historic Data Working Group of the U.S. Federal Geographic Data Committee (2015), was not considered, because at the point of writing this thesis it was still at the draft stage.

These selection and appraisal guidelines with special focus on geodata propose a varied field of criteria and recommendations. It must be said that not all guidelines use the concept of criteria, and the use of checklists is not universal. Some guidelines express formal categorised criteria where others use the same concepts in textual recommendations or in the form of questions. We can identify three different kinds of recommendations in increasing level of applicability/concreteness: guidelines on value, appraisal criteria and indicators. We shall present all extracted recommendations in the following chapters.

3.2.1.1 Types of value

The most general recommendation we can find in the guidelines is to keep records with 'value'. The only documents that do not mention value are NARA's geospatial tips for scheduling geospatial records. All other guides mention value in general or name a specific value as a reason to preserve. The collection development policy for the National Geospatial Digital Archive does not use the term value, but it asks if the records are important for the user, which is a way to express value to the user. The first step in identifying the value in data is showing the types of value that could exist. For example, the Swiss Federal Archives divide values into three groups: legal and economic importance, administrative importance and historical and social importance, although indicators and comments on the individual criteria are short. Their criteria for geodata appraisal are the same as those proved and tested on all archival material. The correct detection of value is, in this case, passed from archivist to archivist through experience. Values can be associated with purposes. A document can have a value in the accomplishment of a certain purpose or can have value to a certain person.

Depending on the moment of appraisal, different criteria can apply. When a first evaluation takes place between the active and semi-active stage of the documents, legal and administrative value might have more weight than when appraising definitive archives. When appraisal is made between the semi-active and definitive archival stage, legal constraints might disappear and historic and scientific value gain weight.

All the values presented in the following chapters were mentioned by at least one of the guidelines. Only one document explicitly defines 'intrinsic value'. For most, no clear definition is given, which is why we extracted a synthesised definition from the context of where the concept is used in various guidelines. Most values are interrelated. Values such as temporary value and permanent value are not retained, because each value can have temporary or permanent characteristics.

3.2.1.1.1 Administrative value

Administrative value refers to the significance geodata have to public administration. It is common practice that legacy data are passed to the archive if they do not serve the purposes they were collected for any more. When legacy geodata lose administrative value to the authority that created them, they can still have business value to other bodies of public administration. Therefore, data should be checked for administrative value not only when appraised between the active and semi-active archival stage but also when passing to the definitive archive, when one might suppose administrative value has expired. When it comes to detecting administrative value, the authority responsible for the geodata is best placed to do so (Ellipse). In the guidelines we find two explicit examples of data sets that are defended for long-term preservation with the argument of administrative value: data that document the placement of critical infrastructure (GeoMAPP) and the coast data set for the state of Maine. Maine argues that the coast map series 'may have future value for reference or restoration of features, as supporting data used in public policy decisions, and in reports, maps, graphs, or tables' (Maine). Ellipse suggests looking at the frequency of demand by public administration and the producing authority. Of course, data requests can only be a reliable indication if the data are only available at a unique distribution point or if all sources where the data are available can be included. When administrative value is so high that the data become vital for the holder, this value can become a selection criterion for prioritisation or additional security or service level (eLegacy).

3.2.1.1.2 Legal value

Geodata have legal value when they can be evidence of rights and duties under national and international law or when they are suitable for use as evidence in legal proceedings (Ellipse). Kentucky suggests that cadastre or land registry data are evidence for rights and duties in the long term and DCC advises that the Kyoto protocol has an influence on the legal value of geodata in general. Even though it is not explicitly said, county borders hold legal value for the California Spatial Information Library. They were selected for long-term retention because they are considered vital for the institution. Like cadastre data, political borders hold information on the geographic extent of rights and duties. If data have legal value to the holder, they also have economic value. Indeed, evidential data can avoid fines, shorten legal processes, or lead to capturing financial benefit by third parties. Legal value is not necessarily tied to the data holder. Public archives assess data value for all citizens. Regarding transparency obligations, legal value can reside in data providing accountability of institutions' actions and decisions. NDSA-LoC explains that accountability for research data is becoming an argument for research data preservation, because it is in line with open access initiatives. Stakeholders in academia that adhere to open access initiatives favour access to scientific information, which eases verification of data and related conclusions.

3.2.1.1.3 Economic value

Ellipse distinguishes between legal and economic value. However, legal value might be translated into a sharp monetary metric. On the one hand, non-compliance with legal obligations can result in fines or have economic consequences. On the other hand, missing evidence of duty of a third party or the right of the claimant can represent opportunity costs.

Conway states that 'commercial value of earth observation data drastically reduces as data become older than three years' (Conway et al., 2013). Shaon seconds that: 'long term benefits are likely to be intangible, so it is advisable to concentrate on short and medium term benefits' (Shaon et al., 2011). Intangible long-term benefits and indirect benefits not perceived by the holder make any legacy data unattractive for archiving. Luckily, archives created as a duty for government do not necessarily have to show a positive return on investment. Nevertheless, especially in times of struggling for budgets, it seems necessary to defend the interest in archiving geospatial assets by making individual business cases. To do so, the cost of elements as much as the benefits have to be monetised. Caruso et al. (Caruso et al., 2013) have found ways to allocate a cost to each preserved data asset. Because a return on investment can hardly be expected, cost-benefit models that include social, environmental, scientific, humanitarian and security value are more appropriate. To monetise these abstract values, some have gone so far as to calculate the annual cost of the worldwide environment (Harris, 2001). The geoarchiving comprehensive cost-benefit analysis guidance elaborated by the GeoMAPP project partners distinguishes between three types of benefits: cost savings, cost avoidance and penalty fees (Geospatial Multistate Archive and Preservation Partnership [GeoMAPP], 2011c). Cost-savings refer to things or actions that cost the company less than before archiving; for example, searching for data. Cost avoidance refers to things or actions that do not have to be paid for thanks to archived geodata, such as regeneration of data sets. Finally, penalty fees refer to income thanks to archived data; for example, fees assessed due to the detection of illegal practice through data analysis.

3.2.1.1.4 Social and cultural value

Some guidelines recommend checking for social value, as for example the potential for conflict of data, media interest and social impact the data have or will have. Ellipse offers the example of keeping data that document the influence of individuals or groups on law, culture and society. GeoMAPP searches more generally for data that documents citizenship. From Kentucky, we learn that data with maximum social value must be ‘useful to characterize the impact on society or to characterize the uncertainty of societal impacts’. When it comes to determining social value, LTDP declares categorically that ‘earth observation data constitute a humankind asset’. This is the point of view of a geodata producer. On the one hand, it is not surprising that it declares its product as important. On the other hand, we should consider that the producer is best situated to judge the social impact of the data and the potential for scientific reuse. The arguments that lead to the conclusion of earth observation data being a ‘humankind asset’ are the uniqueness of time-related measurements and the variety of sciences that can make use of it. With these arguments, we can also categorise aerial photographs as socially valuable. When it comes to detecting social or cultural value several means are proposed. Social media is a tool to detect social impact and controversial topics. Ellipse also suggests looking into media databases. The significance of the data producer or data source should be an indicator of (social) value as well (NDSA-LoC).

3.2.1.1.5 Research value

Most guidelines incentivise appraisers to check if data can be useful to science. In this context, many emphasise research value other than historic research by mentioning the possible scientific disciplines or by separating historic value from research value such as GeoMAPP does:

> *There are several types of value that may be considered: Legal, Evidentiary, (enduring) Historical, Research, Ephemeral, Administrative, Fiscal, Economic (e.g. the cost to replace).*

Duerr proposes to contrast the scientific merit or utility to what it costs to achieve the desired level of service. Unfortunately, the potential scientific merit is difficult to estimate.

'Current use' could be employed as an indicator, but in most cases does not distinguish between research and other uses. Download statistics typically can give no answer to the final use of the downloaded file, whether for scientific, commercial, administrative or leisure purposes. An exception might be the service Edina[30] in the United Kingdom, which is tailored to academic use, and where all access is supposed to be purposed at research and education. NSIDC-NASA intents to measure what it calls 'data activity' by distribution metrics such as data citation in scientific publications, knowing that this is still a not very reliable indicator, as data citation is not yet a generalised habit between academics. Furthermore, NOAA reminds us that

> *[w]hat is of relatively low research use today may become of great research use in the future. Perhaps even more important and difficult to predict are the issues and topics that will be considered of significance in the future. Nevertheless, it is important to consider this question in making appraisal decisions. It is necessary to consider the kinds and extent of current research use and to try to make inferences about anticipated use both by the public and by the Government* (The National Oceanic and Atmospheric Administration [NOAA], 2008, p. xy, citing NARA).

NOAA (again citing NARA) proposes to draw inferences from current to possible future use. Doing so we should base our calculation on stable research interests, as research trends are a poor guide to the selection of records as archives (JISC, 2007). The time span between the detection of the research trend and the point at which legacy data are appraised and readily accessible at the archive might be too long. Ellipse identifies all of the following criteria as shaping benefits for research: interpretation potential (informational value), interconnectability (addressed in chapter 3.2.1.2.4), demand of similar data sets (addressed in chapter 3.2.1.3.1), coherence and continuity with existing records (addressed in chapter 3.2.1.2.3 and 3.2.1.2.4) and diversity of themes and sources. Most of these aspects cannot be assessed by analysing the data set in isolation but must be informed by knowledge about the collections already present in the archive.

[30] https://edina.ac.uk/

3.2.1.1.6 Informational value

Informational value is very similar to administrative value. But most texts distinguish administrative value – given by primary users – from informational value, the value to administrators and members of the public who are not primary users of the data. The term informational value is used by EROS, eLegacy, Ellipse and Maine. ARK refers to the relevance of content of data.[31] Informational value is distinguished from evidential value by eLegacy. If we suppose that data holding evidential value can be used for legal verification, other data might 'only' document important events and decisions but not serve as a legal document. Legislation in Catalonia identifies official geodata as those that can be used for accountability. Other geodata might be authentic and have intact integrity but not be allowed as evidence.

3.2.1.1.7 Intrinsic value

Some guidelines mention the intrinsic value of geodata. We get a hint of what intrinsic value can mean by exclusion of other values such as economic, research, administrative, legal and fiscal value (NSIDC-1). A positive definition of intrinsic value is available in NOAA:

> *Records with intrinsic value are rare and possess one or more specific qualities or characteristics as defined by NARA. These include but are not limited to records in an original form that document an early media type (e.g., glass plate negatives, wax cylinder recordings, etc. – Note that only a representative sample would have intrinsic value and not the entire collection), Aesthetic or artistic quality (e.g., manuscripts; photographs; pencil, ink, or watercolor sketches; maps, etc.), Age (e.g., Generally, records of earlier date are of more significance than records of later date).*

The EROS homepage picks up similar examples and adds to the definition of intrinsic records: 'Records considered "famous", linked with important events, unique, possessing artistic merit, of high monetary value, rare photographs or a unique photographic format'. We recognise in this expression other values such as social, historic and economic value. While the expression 'famous' could be an indication of social or historic value, records linked to important events are definitely of social value. EROS' example mentions the adjective 'rare', while most other guidelines look for uniqueness (eLegacy, LoC, DCC, ARK, GeoMAPP, Maine,

[31] Translated from the original text: 'Inhaltliche Relevanz'.

GeoConnections, NOAA, DANS and NDSA). For more details on uniqueness see chapter 10.1.2.7 about the appraisal criterion ‘uniqueness’. When we understand Intrinsic value as an essential value belonging naturally to the data, as defined in the Oxford dictionary, data cannot lose this value with time. Therefore, EROS counts historic and social value as intrinsic value. Characteristics like data uniqueness, originality or quality, which increase data value, are part of the permanent and therefore intrinsic value.

An indicator for intrinsic value might be the time frame or geographic extent covered by the data. While many guidelines ask for the time frame the data covers (EROS, NSIDC-2, GeoMAPP, NOAA, DANS) only one gives hints on the interpretation of the answer. NOAA citing NARA indicates that ‘Observational records covering a long time period tend to have more value because they enable long-term patterns to be identified and thereby increase confidence in the reliability of data and the conclusions drawn from them’. Additionally, DANS gives the size of a data set as an argument for value. If size is due to the length of the time range covered, NASA’s argument is valid. If the size stems from resolution, detail or vast area coverage, NASA’s explanation can be applied accordingly and would be an indicator of higher potential use.

3.2.1.1.8 Historic value

Maybe most evident when it comes to long-term preservation is the evaluation of historic value. Even though it is mentioned by almost all guidelines, a definition of historic value is seldom given. EROS relates it to intrinsic value. As said in the chapter about research value, it is often distinguished from research value. Ellipse addresses historic value in relation with social and cultural value. On the one hand, we could define historic data as those documenting moments of cultural and social importance that we believe will stay important in the future. In this sense, the historic value is the projection into the future of the social and cultural value. On the other hand, DANS explains that data can be kept for general historical research. In this case, data should allow reconstruction of their production process. As related to research data, DANS phrases it the following way: ‘It is interesting for historians (history of science) and e-scientists to reconstruct the research process. They need the research data to do so, but even more important is contextual information on the origin and background of the project.’ With this definition, historic value can be understood as the capacity to document important changes or actions of its own production process. As compared to the capacity to document natural or artificial change on the earth's surface, which is of interest to a variety of scientists, the documentation of production processes is considered of interest exclusively to historians and the history of science. One criterion related with this value is 'data completeness', especially the presence of sufficient documentation of the context of creation.

3.2.1.1.9 Permanent, archival or continuing value

Documents with permanent, archival or continuing value are those that have at least one of the above-mentioned values for a time frame long enough to archive them. Data with permanent and archival value can still be subject to reassessment in the future.

3.2.1.2 Appraisal criteria

When we dive a little further into the recommendations, we encounter more practical appraisal criteria. In some cases, appraisal criteria are kept sufficiently vast to be applicable to all material. In other cases, special criteria are elaborated for different topics (e.g., only for geographic information) or for different media types (e.g., only for film material). Criteria can be exclusive or inclusive. Exclusive criteria express properties that lead to exclusion from preservation. Inclusive criteria express properties that favour preservation.

With the exception of Ellipse, we have not encountered explicitly exclusive criteria. Mostly the criteria are meant as input to the overall reflection on a preservation decision. Some appraisal instructions express data requirements that favour preservation, but by themselves are not reason enough to preserve. The DANS calls them preconditions and enumerates explicitly the presence of metadata and determined file formats. Even though criteria are more specific than value they are still subjective and interwoven.

Several guidelines (EROS, LoC, ESA-LTDP, NSDA and NGDA-MIT) include the fact that data are at risk in their arguments for selection. Nevertheless, the risk of losing data is only valid if previously value has been detected in the data. Therefore 'data at risk' is a prioritisation criterion and not an appraisal criterion.

3.2.1.2.1 Appraisal criterion: feasibility of preservation

Surprisingly the most recurring general criterion is the feasibility of preserving geodata. This does not mean it is the most important or the first to apply, but might be the most obvious to include. Each archive has to estimate if ingesting and maintaining the collection will in some way exceed the technical, logistical and financial resources of the archive in the long term. Aspects that can influence the feasibility are technical issues of the file format, quality of the data description and responsibilities for the enhancement of quality that can lie at the archive. Weaver, Meier and Duerr refer to feasibility as 'preservation maturity' of the data (Weaver, Meier, & Duerr, 2008).

File format: NSIDC-2 uses the file format as a criterion for evaluating the feasibility of ingest and further preservation. Similarly, NOAA is worried about file format when calculating the effort for turning archived data useable. Indeed, geodata often use proprietary file formats that require specific hardware and software components. An archive would want to migrate those data to standard open formats, which can be a time-consuming process. Archives that assess data from third parties such as DANS can establish preconditions for acceptance, such as determined file formats. An agreed negative or positive list can speed appraisal. The archive will also document the technical specifications of the format to restore rendering and guarantee usability in the future. These technical factors affect the cost of preservation. Of course, cost and other restraints are always contrasted with the financial and other resources available at each archive. No metric in this aspect can be universal. Nevertheless, a strategy to preserve data in non-proprietary and open-source file formats is generally recommended (NGDA-MIT). When file formats are to be maintained, complexity and stability of the structure should be checked. NSDA warns that projects with added functionalities might depend on short-lived proprietary technologies or external resources such as geospatial web services.

The quality of descriptive metadata, as related to feasibility, has directly to do with the designated community the metadata targets. To distribute legacy data in common data networks, they have to comply with the metadata standards of those catalogues and repositories. Enhancing metadata to a certain quality can be the responsibility of the data provider (ideally) or the archive. The degree of metadata completeness determines which designated communities will understand the data. If the data archive cannot obtain the data quality needed by the designated community, it must add the resources for metadata complementation or quality assurance into the feasibility calculation.

Metadata quality: So as not to turn the feasibility criterion into a black and white decision, NSIDC-2, NSIDC-1 and NOAA include levels of service in the decision process. Providing a complete set of standard descriptive metadata can be understood as a higher level of service to the user community than providing only basic metadata for identification such as described in NSIDC-2. Levels of service can also be established for technical aspects such as the file format. Data that are served in the formats optimised for preservation are of a lower level of service than data transformed to a format readily processable by the user community. Because decisions about levels of service and designated community are made by individual archives, no metric is valid in every situation. Nevertheless, acquiring geodata described in a standard metadata format is best practice.

Access: The analysis of current criteria and appraisal questions revealed that access is a topic for discussion as early as the appraisal stage. As we said, the goal of preservation actions is future access to the document. Users demand access and the EROS appraisal questions sustain this idea by asking, for example,

'Q 2.7: How is this collection to be distributed or accessed?'

Nevertheless, in this case we have no indication if ease in distribution and access would favour preservation or if the contrary would favour exclusion of the data from preservation. There is no prioritisation in the question; no hint about if it is inclusive or exclusive. Imagining that the designated community needs highly accessible data and the data are difficult to publish, this question could be an exclusion criterion.

ESA and NOAA data archives seem to appraise without the feasibility criteria. They encourage or expect adhering archives to provide the resources needed to reach data quality at the required level of service.

Cost of preservation: To evaluate the financial feasibility of preservation, all cost factors must be known. To assist identification of cost factors, EROS lists not only time-consuming operations and technical components such as data housing but also possible sources of income. A complete tool for a cost-benefit analysis is provided by the GeoMAPP project (Geospatial Multistate Archive and Preservation Partnership [GeoMAPP], 2011c). The cost-benefit analysis can provide a sharp metric but the components that lead to the metric result are also based on estimations that are mostly difficult to predict. Furthermore, the determining line on what are sustainable costs and what are not depends on the financial capability of each archiving institution. For ongoing collections, considerations of the frequency of ingest must flow into the evaluation of the preservation effort (NSIDC-2).

Cost of archiving is not only used as a selection criterion, but also as an incentive to do appraisal (DANS). Indeed, when institutions appraise and select data, the amount of data finally stored and preserved is smaller and therefore financially more sustainable. Nevertheless, it might be cheaper in the short term for the data producer to add storage space than to invest human resources in appraisal and selection.

3.2.1.2.2 Appraisal criterion: usability

The next most mentioned criterion is usability. Just like feasibility, usability is complex and depends on various aspects. First, content understandability must be guaranteed; second, technical requirements must be complied with; and finally, the access environment and the user's condition and disposition must be in accordance. The criteria that are used together with usability can be grouped into these three categories: data quality in the sense of metadata that determines understandability; technical issues, file formats and the technical knowledge base of the user that determine if he or she can render or run the data; and accessibility, legal issues, authenticity and reliability, which pertain to the context of consultation. The technical aspects can be seen as basic conditions the other two aspects are building on. For example, usability is zero if understandability is increased by adding glossary and legends to the data but there are neither external technical specifications available to render them nor are they part of the designated community's knowledge base. Nevertheless, understandability is equally important. If a data description is not in accordance with the designated community's knowledge base, there is no usability for that specific user community. However, data might still be interpreted by other user groups with advanced data knowledge. As an example of the importance of the third aspect of usability – the access context – when data can be rendered and run in a modern visualization system, metadata is sufficient to understand the intellectual content. But if there is no access right, the data are not useable. In this last example, data can gain usability when access rights restrictions are temporal. To structure this review, we will shed light on each of these characteristics individually. Understandability depends on data quality: presence and completeness of metadata, clarity of explications of measurements and their accuracy and in some cases the presence of other data sets. We discussed how to measure data quality in the chapter about feasibility. The presence of other data sets, however, can be equally important to comprehend a data set. While aerial photographs can be considered self-explanatory, individual thematic data layers might only make sense in the presence of others. This calls for the need to appraise those layers as a whole or to check if they are available and preserved. As an example, the hydrology layer might be used to calculate change in the surface of lakes over time, with no extra layer needed. But when the social impact of a historic event such as the construction of an embankment dam needs to be reconstructed, layers that show what was there before the water are very important. Even if a future researcher understands what the layer represents (say understandability is guaranteed) he or she might not be able to use the layer for research when reference data are missing (usability fails).

Technical issues can reduce usability. File formats or data structures that need uncommon hardware and software not easily available can impede a user from using the data. The technical knowledge of the designated community plays a role, because an understanding of the technology used on the data can open the door to their use if technical requirements are obtainable. Usability studies show that the general public as user does not investigate further if data do not show automatically on screen on demand (Hoa Loranger & Nielsen, 2006).

EROS proposes to look for the availability of training from the data creator or holder. This could be a measure to increase metadata quality or to obtain technical specifications where missing. If training is meant for the archive, it influences the feasibility criterion, because training might speed metadata completion. If training is meant for the user, it is a service level.

Authenticity, reliability and integrity are related to the context of use. Reuse requires authentic and reliable data; for legal or administrative evidence and technical purposes to a higher degree than for leisure. However, leisure users need authentic and reliable data if only because they might have less expertise on how to contrast those data with other information sources. An indicator for authenticity and integrity is the trustworthiness of the data source (producer)and that of all the data holders on the way to the archive, including the final repository. Reliability is given when data are what they seem to be. Checking reliability implies knowing with what goal the data was collected (purpose of the data). Authenticity, reliability and integrity are therefore not exclusively detectable by assessing the data but also the context of their preservation. Information about who the data producer was, to what purpose the producer created the data and who the intermediate data owners were, must therefore be passed on. The ISO 19115 metadata standard suggests capturing information of this kind in the field 'lineage', ergo in EROS, the metadata field 'lineage' is used for assessment of reliability and authenticity. To be maintained, authenticity, reliability and integrity have to be monitored during the entire data life cycle (DCC and GeoConnections). Integrity should be checked regularly. This can be done by a so-called hash that should be created as early as possible for the final data set. A hash is a short control number that is generated by an algorithm using all bits of the data. At determined time intervals or at critical points in the data life cycle, such as during data transfer, the hash is regenerated and compared to the previous. The same algorithm applied on the same data produces an identical control number if all bits and bytes have been transferred correctly and data are unchanged. It is advised to ask data providers for hash numbers so that the archive can compare them after data reception.

A general preservation principle recommends keeping the original. This is true for paper archives where the original is the most faithful transmission of evidence. With digital objects, sometimes it is impossible to extract the very first version in which a file was created, such as in special sensors that capture and output their data in a distinct form. Considerations might be made to keep the earliest 'useable' version available (LoC). NOAA procedures for scientific records appraisal and archive approval integrate the concept of usefulness and define original data as 'data in their most basic useful form' (The National Oceanic and Atmospheric Administration [NOAA], 2008). NSDA-LoC warns that the original as opposed to processed data can lack useful features such as the geographic reference. Due to the many digital versions some data are available in, concerns about usability influence the decision on the version to keep. Some users might wish to keep not only the first useful but also the most useful version. Taking the above example of the aerial photograph again, we would then preserve the orthorectified and georeferenced image and not the first useful output of the sensor. Nevertheless depending on the goal, the sensor output is more useful. For Kentucky, the 'particularly manipulable format' is a rationale for retention.

3.2.1.2.3 Appraisal criterion: organisational focus

Each guideline contains a question on the organisational focus: do the data fit in the collection? The answer will depend on the role of the final repository (archive, library or collection institution). Or is there a legal obligation, a mission statement or an ethical policy that drives preservation? Some guidelines go into detail enumerating the limits of scope: territorial, temporal or thematic limits or interest for a certain population. Most governmental archives give priority to geodata that depict their territory. NGDA-MIT gives details about their collection policy: local geodata are collected in larger scales while country-wide data sets are archived in smaller scales.

The question of whether data have to be kept by law is easy to answer. Legal obligations are public and known by the archives that are concerned by them. An ethical obligation might be reflected in the mission statement, such as providing verification options to research data of a specific thematic field. Wider interpretation is possible when the interest of geodata to a user community must be evaluated and the only incentive is the satisfaction of the user need.

3.2.1.2.4 Appraisal criterion: potential future use

Because reuse is one of the objectives of preservation (LTDP), it is good practice to check if data can and will be reused. However, most guidelines warn us that it is difficult at best to predict future use (Harris, 2001; The National Oceanic and Atmospheric Administration [NOAA], 2008). Data are reused when various conditions co-occur: they present value to some users, and usability and accessibility are guaranteed. Kentucky expresses it this way:

> *Regarding re-use, the value of research data depends largely on various factors such as quality, uniqueness, repeatability, production costs, scholarly use (now and in the future), risk of loss and indications for reuse (in publications or by request)*.

As we explained earlier, both ‘value’ and ‘usability’ are very complex to determine. Some guidelines start by searching for barriers to reuse. They can be of a legal nature, as found in ARK: data protection or blocking periods could overlap with the data age users are interested in. Technical barriers also reduce usability and in consequence reuse, such as if data are served in a file format that needs transformation before it can be processed or needs additional software (for example a GIS) for rendering that is not natively installed on most PCs. EROS suggests verifying if the preservation level (e.g., bit level) is useful for the user community. For further clarifications on barriers and endorsement of usability we refer to chapter 3.2.1.2.2, about usability. Other guidelines advise to start by searching a potential user community. NARA advises in the NOAA document to check if data can be used by the following sectors: legal, science, commercial, education, engineering or resource management. DANS seconds that: 'non-academic researchers', for example journalists, interested amateurs such as local historians, genealogists or others are also increasingly using data.

The most evident type of reuse is change analysis. Types of changes mentioned include geophysical change, growth pattern, key sociological and environmental characteristics and geographic distribution. GeoConnections affirms that 'decisions about our economy, environment and society cannot be based simply on current data; temporal analysis is required to identify trends, evaluate impacts and make informed decisions'. Maine goes a step further and asks why 'tracking periodic changes [is] likely to be important'. Indeed, this reflection might answer the question of whether data will be used for scientific purposes or for more administrative purposes (e.g., urban planning) and can therefore assist with the alignment of data and designated community. The type of change that should be captured might influence the frequency of snapshots taken for the archive, or the retention interval.

The first indicator for this criterion is the temporal range the collection already spans. Because the spatiotemporal range of data collection is a required metadata element in INSPIRE, this information should be easy to obtain. A second indicator is the potential continuation of the collection in the future. Some collections must be offered to the archive when they are still updated, such as layers pertaining to the base map. For a third indicator, the collection should be compared to existing archived data sets, to see if they contain comparable data captured at earlier times. It might be that the appraised data set can be considered the continuation of those archived earlier (GeoMAPP). Accordingly, data that enable interoperability with other data sets or new methods of data collection are of higher potential use. Ellipse calls this characteristic interconnectability of data. Documentation such as reports on changes in collection processes, instruments, etc. can indicate the degree of convenience for comparisons. Geodata might be interconnected with data sets from other domains. EROS asks if the data represent a complete population or are combinable with statistical data of the population. This is especially interesting for research and planning that needs to compare with poll data.

Becoming very concrete, LoC and NSDA-LoC propose three types of data that are of potential interest to wider user communities for science and historical research: orthoimagery, land records and transportation data.

3.2.1.2.5 Appraisal criterion: data quality

Quality can refer to a variety of aspects: data quality as spatial accuracy, resolution and measurement certainty; quality as completeness of descriptive and technical metadata; and material quality. The material quality aspect includes consideration of the type of container the data come in and its physical state, as well as the software and file formats the data depend on. Archives that get offered collections from third parties must appraise those aspects. The physical state of the container is probably the easiest to judge by an archive with experience in digital archiving. Thorough documentation of such characteristics is the key to correct assessment. Spatial accuracy, resolution and measurement certainty can be compared to accepted standards or best practice to judge the quality. Nevertheless, such information must accompany the data in order to be able to compare. Best practice is to let the producer judge measurement quality because of its thematic expertise. Metadata quality has been treated in the chapters about feasibility of preservation and usability. As an indication of what level of documentation is good enough, GeoMAPP suggests that when geodata are well-enough documented for the primary user they are worth long-term preservation. When file formats are known, it should be verified that they correspond to known standards, as one of the quality measures.

For aerial photographs, content quality can apply. One criterion used is cloud coverage of the scanned area. The European Commission Image 2000 Project[32] allows 0%- 5% cloud coverage. Pictures with higher cloud coverage are not stored. This criterion can be effective in an early management state by the data producer. The degree of cloud coverage can be noted in the appropriate metadata field and thus inform the quality evaluation at preservation.

[32] http://image2000.jrc.ec.europa.eu/index.cfm/page/image_selection

3.2.1.2.6 Appraisal criterion: data completeness

We are excluding metadata completeness from data completeness, because we treated metadata quality in the chapter about feasibility of preservation and briefly in the chapter about data quality. Data completeness can imply the presence of auxiliary documents (measurement instructions) and the presence of related data (base maps). Nevertheless, the importance of both of these complementary information types depend on the specific use that will be made of the data. Data completeness in cartography most often means the geographic extent of the data. This should be controlled at production. Missing areas are discovered quickly when compared to a corresponding data set of the same time series. Because of user requirements for completeness NSIDC discussed associating two related data sets to complete spatial coverage (National Snow & Ice Data Center, 2006).

3.2.1.2.7 Appraisal criterion: uniqueness

Eleven out of 18 guidelines recommend checking for data uniqueness. Appraisal questions for public archives are designed for documents that are organic results of administrative processes and therefore unique by nature. Nevertheless, copies can exist and current practice is to eliminate them because they are not original. Geodata in digital form can be copied much easier and existence of data copies can be expected. Apart from data copies, the digital nature of geodata adds new aspects to uniqueness: repeatability and ‘redundancy and substitutability’. Uniqueness in any of these ways favours geodata preservation. All earth observation data are non-repeatable (DANS, DCC), because the time of the observation is an important characteristic of the data. Other process stages such as generalised maps might be repeatable when input data and the generalisation process are preserved.

Often geodata are held at various distribution points (ARK). An archive should check which would be the best provider and avoid ingestion of the same data twice. Legal issues and technological feasibility of data transfer can play a role in the decision. Accordingly, NOAA checks if data exists on other media, eventually even in analogue form. The decision about the preservation term is based on the available media, its feasibility of preservation and the originality of the media.

Geodata might be redundant or substitutable when information is contained in various data sets in different forms. This might be the case for aerial photographs in their various process stages or vector data at different scales. Elimination of derived scales is already current practice in Hessen (Germany) where the smaller scale (1:50,000) that is derived from the larger scale (1:25,000) is eliminated (Frick & Najar, 2009). Frick and Najar argue that smaller scales can be regenerated. Base map information can also be contained in thematic data that was merged with base layers.

The EROS question set also asks for the originality of the data set. Are the data different in some way? DANS lists originality in the factors for valuable data, but does not further explain if originality means 'the original' or 'curious and inventive'.

3.2.1.3 Indicators

Breaking down values and appraisal criteria, we get more practical and applicable indicators. As compared to values and criteria, indicators are measurable or at least produce a clear positive or negative answer when applied to the data. Indicators can inform one or more criteria or value arguments.

3.2.1.3.1 Indicator: current data use

All data that are accessed have a perceived value to the user community. Data use is therefore an indicator of value, albeit not the only one. The criterion is not reversible: when data are not used, this does not mean they have no value, because they can have other types of value or there might be an accessibility barrier.

Current use is a code used for 'perceived current needs', 'known research importance' and the frequency of requests by the science community (ELegacy, LoC/GeoMAPP and Ellipse respectively). Morris believes that if data are used in a broad range of applications this is an indication that they will be relevant in even more unknown applications in the future (LoC). Nevertheless, we have to be careful with declaring currently-not-used data as useless for future users. As mentioned in the chapter about research value, current use is not a reliable criterion to determine future use. In fact, it is not even a criterion but an indicator for the value and usability to certain current user communities. Current use is also a powerful argument to maintain documents. An existing or potential user community is an argument against destruction in the Maine archives. Change in the frequency of use is a trigger for the transmission of archival documents. Decrease in demand indicates the point in time for data to pass to the next archival stage or service level. Indeed, retention schedules and guidelines should be informed by knowledge of current use (LoC).

Availability is a topic when assessing data use. We have to know what is available and where before we can count the use of individual data sets. Therefore, several preservation projects start with inventories of available geodata (North Carolina Geographic Information Coordinating Council, 2008) or recommend inventorying or use of existing inventories and catalogues (NSDA-LoC, GeoConnections, GeoMAPP). Data sets that are widely available are already more predetermined to wider use. When it comes to measuring the use of individual data sets, we must distinguish between online and offline data. Online data can be controlled by a system that records access and download statistics. Download statistics cannot be used in isolation to measure current use, as more and more data is served in flexible online views. Access clicks by online viewers who combine several data sets into a renderable map do not inform about which data were really needed. For example, a user might view a map with several features but only be interested in the transportation system. Where data are not publicly accessible, the data provider should deliver information about all bodies the data were offered to and are known users of the data (EROS). Information about all the places data are held might help quantify current use, highlight types of users, indicate possible redundant data holdings and point to derivative data and user groups that might depend on the appraised data set. Similarly, the data provider or data holder should record all data demands from the provider side, public administration, representatives of research, companies or private individuals. This allows for counting demand and detecting change in demand.

3.2.1.3.2 Indicator: legal barriers

Legal barriers are negative indicators for the feasibility of preservation, accessibility and, in consequence, usability. Legal issues include such issues as: the data producer does not want to grant open access to data because of intellectual property rights; the data contain sensitive information or provenance of certain layers is unknown. Not only content but also databases and data structures can be protected by copyright. Legal barriers are easy to detect when legal issues such as copyright and provenance of data layers are properly documented or discussed in the transfer contract. Luckily, most legal barriers are not permanent. Even sensitive data become publicly accessible in the long term. Nevertheless, the archive should check if the data are still of interest to the designated community after sunset of the legal barrier. The copyright issue is sometimes not a black and white question. If data can be handled in a digital rights management system, the data producer might agree to the transfer to the archive. A user willing to pay for access can then still use the data. Nevertheless, most likely the archive will want to include feasibility considerations in the evaluation of this measure.

3.2.1.3.3 Indicator: standard files and metadata

Metadata organised in standard ways is easier and faster to check for completeness than metadata that is not so organised. Information is easy to extract when the standard is well known and might even be automated. Therefore, it is a factor in the cost calculations for ingest, quality checks and eventually for the appraisal itself, as a time-saving element. As a consequence, it is a feasibility indicator.

3.2.1.3.4 Indicator: data source

The data producer itself can be a source of confidence and a sign of quality. In the case of research data, the reputation of the data producer can also be an indicator for the social and research value of the data sets (DANS). The data source might determine whether there is a legal obligation to preserve or more generally indicate whether the data correspond to the organisational focus of the archive. On the one hand, NGDA-MIT states explicitly that all data sources, private and public, can provide data to the repository. On the other hand, public archives might accept only government records and might not necessarily preserve data generated by any other agency or person.

Approved lists of data sources, accepted file formats, metadata standards and legal licenses might increase the utility of these indicators.

3.2.1.4 Influencing factors: Levels of service and accessibility

Accessibility indicates how simple it is for users to access the data. Accessibility is a criterion that influences usability. Apart from the legal restrictions that impede access completely, there are conditions that only partly influence accessibility and in consequence usability.

Among the latter are the availability of data (online or dark archive, the width of distribution) and the level of awareness by the user community about the data's existence. Accessibility is higher when data are discoverable through different channels and well described (quality descriptive metadata). When we remember EROS' appraisal question about access, we can assume that convenient distribution and access would favour preservation, because a heavily accessed data set is of user interest. In another scenario, data that are heavily accessed might not be selected for preservation because they might stay accessible at the producer site, or archiving might be postponed to a later point in time. Heavy demand for access is related to cost and therefore might be dissuasive to the archive. If users pay for access, frequent use might dissuade a producer from sending data to the archive. We see that in this case there is no tendency for the question proposed by EROS to be an inclusive or exclusive criterion. Similarly, ARK includes a question about where data are easier to access in its appraisal checklist. The degree of accessibility offered for a data set is a service level. NSIDC-2 proposes to compare each data set to others already accessible to the public, to estimate the access level demanded and in consequence the support level needed. This information in turn is an input for the feasibility criterion. Accessibility is related to availability. In the perception of the user, when data are barely accessible they are perceived as unavailable. If data are available at different location, ARK suggests asking where they are more accessible.

Accountability: while for research, accountability is an incentive for including research data in an archive, in the case of government archives such as in Canada (GeoConnections), accountability of all documents generated in the organic process is a general requirement for data management.

3.2.2 Apart from appraisal: achieving smaller collections

Ideally, an archive would define objective thinning criteria to be applied to a collection before appraisal. Thinning criteria are considered objective when they can be applied without other considerations. The thinning criteria evaluated by Project Ellipse are redundancy, duplication, scale, and acquisition interval.

Redundancy: Redundant data should not be preserved twice. Data are redundant when a smaller part of data is contained completely in another data set. This might be the case for generalised maps that stem from larger scales or thematic data that contains base data layers.

Duplication: Copies of data should also be eliminated. When a data set exists twice, for example in two file formats, it might be unnecessary to keep both versions. The question here is to determine if two versions are truly identical. This is often not the case, as different file formats allow different services and desirable properties might exist only in one version. Some archives have strict requirements when it comes to file formats. Data that do not conform to required file formats are not ingested. Data offered in those formats receive negative appraisal decisions or must be transformed into one of the allowed formats.

Scale: Maps of different scales from separated data capture processes (where one scale is not derived from the other) can still be considered as partly redundant if they depict the same features. Certain research questions can be answered by either of the map scales.

Acquisition interval: The larger geodata acquisition intervals are chosen, the smaller will be the final collection and the lower the time resolution. In contrast, a higher frequency of acquisition will result in large collections.

3.2.2.1 Selecting data objects of the production chain

Data objects with the same intellectual content at different points in the production chain of spatial data might come close to being copies or redundant. Therefore, considerations about which level of processing must be preserved can occur. LDTP urges its members to preserve the following data sets: ‘Raw data, Level 0 data and higher-level products, browses, auxiliary and ancillary data, calibration and validation data sets, and metadata’. Level 0 is the rawest form of data enabling reprocessing (Harris, 2001) and according to Harris is the preferred level for long-term preservation for space data. NOAA also advises that ‘it may be warranted to appraise as permanent both a raw version and one or more processed versions of certain data’. In the eyes of the Library of Congress, the rendered map might not be considered as redundant to the individual layers. LoC explains:

> *The true counterpart to the paper map is not so much the dataset as it is usually a combination of datasets, which have been synthesized and displayed in a particular*

> *manner. These derivative products represent an entirely different information object, the capture of which may stand as an additional objective, which does not preclude or take the place of capture of the actual datasets (and other technology components) needed to create these outputs*.

As practical advice, making decisions about data objects can be started by identifying the process stage of the present data and locating possible other existing processing stages (EROS, ARK, NOAA, DANS). In relation to research data, DANS distinguishes between primary and secondary data. 'Primary data are data in their most basic and elementary form: unembellished, pure observations. These are often the raw data, i.e., data not yet influenced or edited by researchers. Another distinction can be made between data that have been published or been the base for publication by others. It should be noted that DANS does not give direction on which data to prioritise. 'To limit the amount of data, for example, a decision might be taken to preserve only the primary data or only the secondary, published data.' NSDA reminds us that derivative information objects can have added value and be created by other producers as the primary data. NOAA relativizes that statement, accrediting derivate products that can be regenerated on demand as having inferior archival value.

3.2.2.2 Acquisition interval

The frequency of acquisition only has to be determined for data sets for which value has already been detected. It is recommended that archives consider the frequency with which a data set is updated. As mentioned before, frequency must be chosen very carefully with data sets that are interdependent. Decisions about frequency of capture for reference data sets should be decided with awareness of consequences for accuracy and potential accountability of related data. When historic versioning (historicizing) of the whole geospatial database is not an option, snapshots are taken. Ellipse proposes three models for combining the frequency of capturing reference data with thematic related data. The models have a direct effect on data redundancy (Projektteam Ellipse, 2012). Redundancy and the storage space occupied in consequence should be taken into account in the evaluation of preservation feasibility.

4 The case study: the ICGC

4.1 The ICGC's environment

The ICGC is a public entity that provides services to businesses and citizens. Since its inception in 1982, it has produced official maps for the autonomous region of Catalonia, first in print and since 2005 fully digitally. The ICGC preserves all of its data because they are an integral part of the provided products and services.

The ICGC has never deleted information of its own production but today is concerned about the conservation of this large volume of stored data. On one hand its data lacks proper long-term preservation management, on the other hand, Article 48 of Law 16/2005 orders it to preserve the cartographic heritage of Catalonia of which their maps are part, which means that long-term preservation is not a mandate but a legal mission.

4.1.1 Legal context of digital preservation in Catalonia

In this chapter, we analyse what parts of geodata produced by the ICGC should be preserved by law. Legally controlled preservation of geodata depends on the definition of the term 'document', the interpretation of 'creative work' and the meaning of 'published'. Legal digital preservation is basically determined by the following two laws and one additional regulation:

- Law on archives and document management (*Ley 10/2001, de 13 de julio, de Archivos y Documentos*)
- Law on legal deposit (*Ley 23/2011, de 29 de julio, de depósito legal*)
- Royal regulation on legal deposit (*Reial decreto 635/2015, de 10 de Julio por el que se regula el depósito legal de las publicaciones en línea*), which names the institutions that assume the responsibility of archiving online publications.

As a general distinction, electronic administrative documents should be preserved by the public archives and digital publications by the deposit library. The public administration work process generates documents. In most current administrative processes, documents about the same affair form a dossier. The term 'document' in the law on archives and document management refers to the definition of document in the law of cultural heritage of Catalonia. There, a document is defined as all expressions in oral or written language, in images or sounds, natural language or codified and contained in any kind of material medium and any other graphic expression that is testimony to the social functions and activities of humanity and human groups, with exception of research work and creative work. Preservation, management and access to documents in the above sense are regulated under the law on archives and document management of Catalonia (Catalonia, 2001). Even though the ICGC has to some point an independent structure, it is created by law as a public service and therefore subject to law on public administration. The law forbids public administration bodies to destroy documents unilaterally. It obliges them to hand archives over to the evaluation process and in case of positive appraisal to the final preservation institution.

The preservation of published documentation is regulated by the law on legal deposit in Catalonia (Spain, 2011). All published sources of information are covered by this law, independent of the medium they are published on, which means that it includes electronic publications. Explicitly mentioned are edited photographs, maps, plans and atlases and websites that can be affixed or harvested[33]. The law and its corresponding norm determine that published documents should be handed over to, in the case of tangible documents, or captured by, in the case of non-tangible documents, the National Library of Catalonia. One could argue that orthophotographs are edited photographs, but if they are available for the public on the ICGC's website, they are considered published. The whole website could also be preserved in the category websites that can be harvested. Any document, though, that is a company internal publication or is susceptible to being preserved by the law of archives and documents is excluded from legal deposit. The ICGC holds many similar versions of the same geographic information. Not all of them are published. Legal deposit law does not reach out to the institutions' internal information. In the case of the ICGC's aerial

[33] The law referes to 'sitios web fijables o registrables'.

photographs, the originals that, as a principle in preservation should be given priority, are not published and therefore not preserved by legal deposit. Nevertheless, as graphic expressions and testimony to the social function of the ICGC, aerial photographs are considered documents and should be handed over to the evaluation process of the public archives.

When it comes to databases and online services, the distinction between what is intermediate documentation and what is published is blurred. The fact that map data are created and served in different technological environments makes the case more complicated. Online services that make vector data consultable over the internet require transforming the data from their native GIS environment into a format compatible with the web service. Morris argues that the representation of this data as consultable maps contains other significant information not present in the original database and should therefore be regarded as two different informational objects (S. Morris, 2013). The legal deposit institution in Catalonia considers that databases that can render information they hold visually are creative works. On one hand, this definition makes those databases exempt from the law on archives and document management and exempt from evaluation and preservation by the public archives. On the other hand, unpublished databases such as the original vector information in the GIS are not subject to legal deposit either. As a conclusion, if we look at the vector geodata in its different expressions as separate documents, such as recommended by Morris, the original would not be preserved under either of the examined legislations. As long as vector databases are not rendered directly on the web several expressions will exist and the original version of the vector map as created in the GIS is not preserved by law. Nevertheless, if we look at the different expressions of the vector data as pertaining to one single creative work, we could argue in favour of also depositing the original vector information to the national library, because it is the original of a published work. Other unpublished geodata, such as databases, that come from government institutions and are not saved as ‘creative works’ fall under the definition of document and are regulated by the law on archives and document management.

Under two other considerations the original database of geographic information could be preserved by law. Both are much more open to interpretation: First, the law on archives and document management mentions the role of the preserved documents as forming the

Catalan documentary heritage and in consequence the national memory. To be a complete national memory, documents should reflect all decisions and actuations of a public institution. If preserving a specific technological solution of the vector data is essential to the understanding of the decisions taken in the ICGC, it has to be preserved. However, one snapshot of every technological solution would be sufficient to comply with this criterion and not all historic stages would need to be preserved. Preservation decisions based on this interpretation stand on an unstable foundation. The second law that suggests preservation of unpublished geodata is more specific: law 16/2005 on geographical information and the Cartographic Institute of Catalonia assigns to the ICGC map library the responsibility to preserve the cartographic heritage of Catalonia (Catalonia, 2006). Article 48 specifies that the library should collect, preserve and provide access to cartographic documents produced by the ICGC. The definition of 'cartographic documentation' in article 2 of the same law explicitly includes databases, archives and collections that have anything to do with geography. This definition defends the right of vector geodata in its raw form to be preserved by the map library of the ICGC. Additionally, this law promotes that the map library should receive a copy of every cartographic product that is deposited at the National Library of Catalonia by legal deposit. This gives the ICGC the authority and obligation to preserve its own digital production and makes it the only archive for 'unpublished' vector geodatabases. Both expressions— 'cartographic heritage of Catalonia' and 'cartographic documents produced by the ICGC'—give the map library a clear focus on the region of Catalonia. Because this mission statement is defined by law it will later not be necessary to define a selection criterion that specifies the Catalonian provenance or coverage of documents. Nevertheless, if the library wishes to preserve documents that cover further areas it might give itself a selection criterion about territorial coverage.

4.1.2 History and external context of the ICGC

The *Institut Cartogràfic de Catalunya* (ICC) was created by law 11/1982 of October 8 (Catalonia, 1982). The currently valid abbreviation is ICGC. Article one says that the cartographic institution is commercially, industrially and financially autonomous and assigned to the Department of Territorial Policy and Public Work (*Departament de Política Territorial i Obres Públiques*). Article two insists that the ICGC has its own legal nature and administrative and financial autonomy. Its main tasks, as called for in the law, were to produce and disseminate base maps for the autonomous region, to coordinate cartographic work at the level of the local institutions and to align its work with partners of analogue scope. It also was put in charge of the already existing map library of Catalonia, opening up the possibility to share this responsibility with other public institutions. Finally, the ICGC was tasked to initiate a database of geographic information, which means that the ICGC has been creating digital data since its beginning.

On June 11, 1997, a new law turned the nature of the ICGC into a public law entity that adjusted its activity to private law to increase its efficiency and flexibility exercising the functions it was entrusted with.

It was temporarily merged with the Institute of Geology, then they became two separate entities again and in 2014 they merged once more to become the *Institut Cartogràfic i Geologic de Catalunya* (ICGC).

At the national level, the Spanish National Geographic Institute (*Instituto Geográfico Nacional – IGN –*) performs comparable mapping tasks for Spain. The IGN also creates base maps for the region of Catalonia and is therefore in competition with the ICGC when it comes to some map scales. There are other autonomous states in Spain with their own cartographic institutions that operate at the same level as the ICGC.

At a local level, the ICGC can count on the collaboration of several townships, the Government of Catalonia (*Generalitat*) and groupings of municipalities (*Mancomunitats*) that participate in various degrees. Townships and other political bodies might participate economically in the production of map scales larger than 1:5,000. Co-funding contracts regulate the ownership and user rights of the data. Furthermore, local, provincial and metropolitan governments collaborate with the ICGC in updating the common Street Database of Catalonia by exchanging data or information about changes in the road network. On an institutional level, the municipalities are represented in all technical commissions and management organs of the ICGC. Finally, on the technical level they take part in the elaboration of technical norms and specifications for map scale 1:1,000, the street database and the local geodesic network.

4.1.3 Internal structure

When the ICGC took its current form in 2014, the institution gave itself a contractual program that is structured into seven areas: three of which are directly related to the basic intellectual field it is responsible for (basic geoinformation, geology and geodesy). The remaining four are transversal to all knowledge disciplines: geogovernment, geodiffusion, geotechnologies and investments. These seven areas are divided into 13 departments as shown in the following table:

Thematic areas			Transversal areas			
1. Basic geoinformation	2. Geology	3. Geodesy	4. Geogovernment	5. Geodiffusion	6. Geotechnology	7. Investments
Urban systems	Geological infrastructure	Geodesy infrastructure	Coordination and legal affairs	Data	Technological development	Investments
Territorial systems	Risks			Tools		
	Geological resources			Services		
				Knowledge		

Table 2: Structure of the seven areas and thirteen departments of the ICGC

Hereafter we briefly explain the functions of those departments that are related to this thesis, excluding the fields of geology or geodesy. Urban systems and territorial systems are relevant because they are the real producers of aerial photographs and map data. Coordination and legal affairs manages, among other things, the registry of official maps. The registry is a deposit of map data that has been defined in the Catalonian mapping plan to be of legal reference. Part of the same department is the Coordinative Commission of Cartography in Catalonia (C4). The objective of this commission is to coordinate cartographic efforts of the autonomous state between all producers. This commission, through its technical task forces, advances the deployment of the Cartographic Plan of Catalonia (*Plan Cartografic de Catalunya*) and the implementation of the INSPIRE directive. As a side benefit, current and legacy data become more interoperable and in the end increase in usability. The Data department makes data accessible through the web and the library. The cartographic library that is attached to this department has the legal mandate to preserve the cartographic documentation of Catalonia (Catalonia, 2006). It is involved in preservation of historic maps, it digitizes them, it adds georeferences and it offers several services to access and download them through the Internet. Its website received 114,230 visitors in 2015; almost every other visitor downloaded a file.

4.1.4 Production and services

Information collection for this thesis occurred prior to the merger between the Cartographic and the Geologic branch of the ICGC. For that reason, and because of the focus on digital data resulting from map production, this chapter only discusses products directly related to the cartographic branch of the ICGC. Aerial photographs and map production can be grouped into three areas:

- Base maps
- Aerial photography
- Thematic maps and data

The ICGC services discussed have to do with data distribution to the general public.

4.1.4.1 Base maps

The ICGC produces three scales of base maps: 1:1,000, which covers urban zones, and 1:5,000 and 1:25,000, which cover all Catalonia. Base maps are created by photogrammetric methods for restitution of the three-dimensional position of objects. When it comes to the 1:1,000 scale, the ICGC has updated 194 municipalities or 37,169 hectares in 2015. This represents 20 % of the municipalities.

The star product of the ICGC is the base map at the 1:5,000 scale, which is the most detailed map covering the whole territory of Catalonia. It has a precision of 1 meter in planimetry and 1.5 meter in altitude. This map scale is composed of 4,274 sheets of which 561 sheets have been updated in 2015. This represents about 13% of the area. The ICGC has updated all of its maps at this scale at least twice. Where urban infrastructure has changed faster, map sheets are in their fourth, fifth, sixth or even seventh edition. Lately the data model changed and all represented objects received a unique identifier. This new data model turns the map sheets superfluous; updating can take place on individual objects independent of sheets or the layers they are in. The 1:5,000 scale exists in two versions. One is registered as a base map and one is derived automatically as a conventional map and visualized by the online application. In the case of the 1:5,000 scale, the two versions carry the same information. In his history the ICGC undertook a major change to the 1:5,000 scale data model that made the data captured previous to the change useless to the new data model. The update process that started in 1996 included new capture of the data, made them GIS compatible and allowed automatic generation of 2,5D views of topographic objects (Pla & Lleopart, 2010).

The 1:10,000 scale is derived from the 1:5,000 scale. To obtain this scale requires computing time executing algorithms to simplify and shrink the vector data and human effort when the algorithms do not fix it. This series is younger and its sheets reach from first to fifth edition.

The smallest scale that counts as a base map is in the proportion 1:25,000. This map was first generated from the 1:5,000 scale by generalization with additional human input, and in the following revisions received input directly from photometric analysis. The 1:25,000 scale map is created out of the topographic base of the same scale. Both versions have parts that have been updated between one and four times. Compared to the base, this map holds additional information that comes from topic databases, such as administrative boundaries, protected areas, soil use and surface type, points and structures of touristic interest and walking paths proposed by the local bodies responsible for the natural environment. Additionally, the street network of the map version incorporates a classification that allows representing street width and pavement. Unlike the larger scales, the 1:25,000 scale map is conceived for printing and its sheets overlap. The series contains 77 sheets of irregular coverage, chosen to suit the city or natural environment it depicts.

Currently the database has no link between the same object represented in different scales. Multi-resolution data models that would be able to maintain the link between an object in different resolutions are being studied. Web services that can generalize large scale data on the fly fully automatically are not in place yet (M. Pla, personal communication, April 3, 2017).

By law, the ICGC is not allowed to produce base map series of the 1:25,000 scale and smaller (Spain, 2010). Those are produced by the National Geographic Institute. Nevertheless, the ICGC can produce thematic products with small scales when necessary in the exercise of its authority. This is why the 1:25,000 scale has irregular sheets and is centred around a topic.

One of these smaller scale series is the series of 1:250,000 scale strategic maps. There are 20 such sheets of various topics. The topographic version is at its 11th edition whereas the transportation map is at the 10th edition.

4.1.4.2 Aerial photographs

The southern and more populated parts of Catalonia are sensed by aerial photographs at a resolution of 22.5 cm per pixel. The Pyrenean area is sensed at the lower resolution of 45 cm per pixel. After orthorectification the images reach a resolution of 25 cm per pixel; 50 cm per pixel in the Pyrenees. The orthophoto is composed of 4,275 sheets and distributed in colour and infrared. The higher resolution is also generalized and distributed to a homogenous 50 cm per pixel over the whole area of Catalonia. These two resolutions build one series because together they cover all of the territory of Catalonia. A second series consists of the generalized orthophotos at a resolution of 2.5 meters per pixel corresponding to the 1:25,000 map scale. This series is composed of 305 sheets. In 2015, the whole series was updated with the flight of the year before. In 2010 and 2011, the coastal zone was sensed with a very high resolution at 9 cm per pixel, which resulted in orthoimages at a resolution of 10 cm per pixel. In 2015, all sheets of this series were processed and now number 747.

Furthermore, the ICGC has a copy of the historic flights of the years 1945-1946 and 1956-1957, the so called ‘American flight – series A’ and ‘series B’ respectively. In 2011 series B was orthorectified and made available in 2015, followed by series A. Both series were made by the United States Air Force under an agreement with the Spanish Air Force (CECAF). A lot of human effort is required to identify points of reference with sufficient accuracy in the changed landscape. Additional difficulty is introduced by the state of art in technology that was used at the time to create the photographs and by the physical condition of the originals.

4.1.4.3 Thematic maps and data

In addition to base topography and aerial photographs, the ICGC produces many more data products. Some are available as layers through the map viewer on the Internet. Here are just some examples: mountain tops, triangulation points, protected natural environments, etc. The various sensors at the ICGC collect data for different economic sectors: the hyperspectral sensor creates raster images that help agriculture to decide whether fertilizer is needed or not on the fields, radar data are used to study movement of land masses in areas where differences from one year to the other can be expected, such as the coastal zone, and Lidar data serves for calculating contour lines, topographic profiles and the digital terrain model. Furthermore, the soil map series classifies soil use into 40 categories and is represented in a colour raster image. Another geodata product is the dictionary of place names in the Catalan language, where every denomination is localized by its geographic coordinates. Finally, the ICGC vectorizes the boundaries of forest fires for the Catalonian Department of agriculture and natural environment.[34]

4.1.4.4 Services

Since 2002 the ICGC has allowed open access to most of its data. In 2015, this practice was generalized so that all products of the ICGC are now licenced under Creative Commons CC BY version 4.0. This means that anyone can use the data free of charge as long as he or she credits the creatorship of the ICGC. The web services that enable users to visualize and download current data on the internet are regularly enriched with additional data and new functionalities are added. A tool allows selecting an area of interest to download various tiles (individual parts of represented geodata of the same scale) at a time. Through the use of another tool, users can calculate elevation and distances of hand-drawn routes over the map.

The prototype Instamap is directed at users without GIS who want to create and share their own layers and represent it over the ICGC information. It is a set of tools for public administration and the general public with which users can visualize and explore information added by others. An additional web service visualizes the historic aerial photographs of 1945 or 1956 over the recent map on a restricted area around the curser.

To increase accessibility to old maps, the map library digitizes not only their own historic maps but rents the scanner to other institutions. In agreement with the owner, the library makes a copy of these additional maps accessible.

[34] Departament d’Agricultura, Ramaderia, Pesca i Alimentació

4.1.5 Production chain of topographic maps (from raw data to the end product)

The production of topographic maps at the ICGC starts with aerial photographs, goes through orthorectification of photographs and ends with different products, such as the geographic database and the rasterized maps for the web services. Morris considers that the end product and the services through which this end product is available are two different production stages (S. P. Morris, 2010); for the services usually some transformation is necessary. Throughout the production of base maps much intermediate data – data not considered by the producer as an end product – is created Data can be intermediate in one process but not in another. This would be the case for example for the orthophoto. In the process of updating the topographic database, the orthophoto is intermediate data. Nevertheless, it is also treated as an end product and distributed as such. This means that whether or not data are intermediate depends on the process. Intermediate data can be calibrated, corrected, modelled, processed, represented, rasterized, generalized, etc. This list of attributes is non-exhaustive and purely illustrative of the various forms data can take. In addition, these attributes are not exclusive for intermediate data, as final products can also be generalized or rasterized, etc.

Here we explain in more detail how the orthophoto is made at the ICGC and which data are created and preserved. At the beginning of the process is a camera with several lenses that captures pictures of different colour channels and infrared in a proprietary format. The process of extracting these images from the camera can be understood as the developing of a digital negative. The negative is in an exploitable format and consists of four files and a thumbnail (see table below).

Suffix	Description	Resolution and bit level
CON.TIF	Thumbnail of the image	8 bits
CIR.TIF	Contains infrared, red and green	High resolution, 12 bits
RGB.TIF	Contains red, green and blue	High resolution, 12 bits
LR4.TIF	Contains red, green, blue and infrared	Low resolution, 12 bits
PAN.TIF	Panchromatic image (tChannel)	High resolution, 12 bits

Table 3: The five files of the digital negative of an aerial photograph.

The CIR.TIF and the RGB.TIF are obtained from the LR4.TIF and PAN.TIF using the camera producer's proprietary process of pansharpening. The ICGC's experience shows that pansharpening with other known algorithms does not provide results of equal quality. Therefore, all files are preserved even though CIR.TIF and RGB.TIF contain the same information as LR4.TIF and PAN.TIF. The digital negative is preserved in a dark archive. From this negative, a master file RGBN with 4 channels at 8 bits per channel is created for further processing.

Due to the curved nature of the lens and the perspective view, all photographs are distorted; only the middle of the photograph shows the exact location and shape of the objects. Therefore, the next step is orthorectification of the photographs, which restores the true distances between objects and the accuracy of the orthophotomap. Orthorectification is a series of processes. First, with the help of control points for which the exact position is known on earth and on the photo, the geographic orientation of the image is adjusted. This is called aerotriangulation. Through aerotriangulation, the x and y coordinates of the central pixel and the rotation angle of the camera in the moment of capture are determined. The second step, obtaining altitude (z-coordinates), requires a digital terrain model. The coordinates and rotation serve as a georeference of the picture. Third, orthoprojection transforms the pixels of irregular scale (due to their position, the inclination of the terrain and the angle of capture) into a pixel of regular size. An algorithm defines the new radiometric value of the pixel. Interpolation eliminates the effects of misalignment of the pixel grids due to the new orientation of the image. After orthorectification, the radiometric values are adjusted within one shot and between shots to obtain a homogeneous image. For adjustment, overlapping areas of shots and a continuous low-resolution image are needed. Next, artefacts such as shadows and stretched pixels are taken care of by carefully choosing the seamlines where adjacent images should be joined. Finally, the orthophotos are generalized to the resolution of 50 cm per pixel. The resulting master orthophoto at the 1:5,000 scale is composed of the RGB.TIF and the CIR.TIF. For distribution, MrSID and eventually ECW and JP2 file formats are created.

Various inputs help to produce the map: automatic change detection between the last and the newest edition of the orthophoto helps to detect areas where updates are needed. Overlapping areas of the most recent orthophotos are used to create a stereoscopic view of the landscape. Human labour is needed to interpret the view and translate it into vector features. Sometimes additional input from field work is needed. Because the ICGC works with tiles, updates of vector data are done for all features in the tile before generating a new version. When all features are captured and laid in place, the vector map is rasterized to JPG or other formats necessary to serve online viewing. The TIFF and PDF formats are served for download requests. The JPG format is currently a popular format used for online viewing, not only for the map, but also for the copy of the orthophotos. Newer viewers allow vector tiles such as the prototype for the 1:1,000 scale.

4.1.6 Production chain of thematic data

Thematic data comes from a variety of sources. Certain subjects require going out in the field to ask people questions or to make personal observations or measurements in situ. Other information can be extracted by scanning aerial photographs by eye. Certain data might be automatically generated from image analysis or captured by sensors. Whether or not information is considered part of the base map or is considered thematic data depends on intellectual decisions and is officially regulated. Regulations and semantics vary from region to region. For example, political boundaries can be considered part of the base map in one country but not in another; archaeological sites might be in the base map when they are considered interesting for tourists or part of a thematic map about archaeology. Any subject with a location can be a theme for a map. Because of the variety of instruments and collection processes, we cannot go into detail. The goal of the production is the same for all data: integrate them into the GIS either as table, raster or vector objects. Most of the data are processed for online display and/or for print products such as atlases.

4.2 State of digital preservation at the ICGC

In this chapter, we give an overview of the results obtained by auditing digital preservation processes at the ICGC with the international standard ISO 16363. The results of the audit should be viewed in the context that the ICGC did not have digital preservation implemented in its management of IT and mission statement. The fact that the ICGC agreed to be audited by a standard with higher requirements than those it had implemented at the time is remarkable and shows its interest in improving its digital preservation services. The audit was administered at the beginning of the thesis and the results were five years old at the time of writing, which means that the ICGC might have changed its procedures meanwhile.

The audit criteria are grouped in three topics: 'organizational infrastructure', 'management of digital objects' and 'infrastructure and risk management'. These three topics come from sections 3, 4 and 5 of the standard. The numbering in the analysis follows that of the standard and therefore starts with section 3.

Following is a summary of the results for each criteria or sub-criteria. The reference to the specific audit criteria is given in parenthesis. The complete list of criteria with their definition can be viewed in ISO standard 16363.

4.2.1 Section 3: Organizational infrastructure

3.1. Governance and organizational viability

We consider that law 16/2005 determines the mission of the ICGC whereas the obligation to preserve the map heritage of Catalonia is part of (3.1.1). Nevertheless, this obligation is not translated into concrete written internal instructions of the institution. For example, the strategic plan of the ICGC preservation does not mention or plan for how to transfer data to another institution in case of complete operational stop of the ICGC (3.1.2.1). Therefore, survival of the ICGC's data is not guaranteed.

There is no written document such as a collection policy that defines the type of documents to be kept by the ICGC to distinguish them from holdings of other institutions or repositories (3.1.3). Lack of coordination at this level can result in doubling financial and time investments between the ICGC and other institutions.

3.2. Organizational structure and staffing

The ICGC has obtained very good results under this criterion.

3.3. Procedural accountability and preservation policy framework

The ICGC has defined the user types its website should serve and identifies the internal producing units as stakeholders of the active archives. ISO 16363 also asks for defining a designated community interested in data to preserve in the long term (3.3.1), which is missing at the ICGC.

The lack of a strategic preservation plan results in missing concrete preservation policies that would guarantee the compliance of the preservation plan (3.3.2). Compared to the requirements of ISO 16363, back up policies of the ICGC are not sufficient and use of standards should be part of preservation policies. The ICGC has mechanisms to check and update its other policies. The same mechanisms could be used to audit future preservation policies (3.3.2.1).

If institutional knowledge depends on the knowledge of an individual, it is considered a risk. The ICGC cannot guarantee that, in case of absence of a staff member, all necessary knowledge for preservation actions would be maintained (3.3.3). Nevertheless, efforts in registering personal knowledge have been advanced through an institutional wiki and through the gazetteer.

The ICGC currently lacks a preservation audit plan at both the internal (self-assessment) and external (certification) level as required by criteria 3.3.6. Execution of the present audit does not suppose the existence of a plan for future regular audits.

3.4. Financial sustainability

The ICGC will be able to comply with the audit requirements as long as it continues to capture sufficient economic funds to alleviate digital preservation risks (for example, for format migration and renewing of the information technology infrastructure).

3.5. Contracts, licences and liabilities

Criteria under this point are not applicable at this time, because the ICGC only stores its own production. In case archiving of foreign data to which the ICGC does not have full rights should occur in the future, this point should be re-evaluated.

4.2.2 Section 4: Digital object management

4.1. Ingest: acquisition of content

Data producers and the IT department that stores their data do not exchange the significant properties of the preserved data (4.1.1.1). Due to this lack of communication, the ICGC risks altering these significant properties against the wishes of data creators when certain preservation actions are performed (such as format migration).

Content acquisition is initiated by the creator. The transfer process from the creator to the IT department assures that all files that were required for copy are correctly stored. Nevertheless, archive managers do not control which content is preserved for the long term. This processing has the following consequences: a) file formats of the stored files are not identified (4.1.3), therefore the ICGC cannot implement a process to supervise the obsolescence of the file formats; b) important content might be missing (4.1.5).

The platform for file exchange for final storage lacks a mechanism to check file integrity (4.1.8).

4.2. Ingest: Creation of the archival information packages (AIP)

The ICGC does not create AIPs, as defined by ISO 16363 (4.2.1.2). There is no normalization of file formats or aggregation of metadata at ingest. For the use of this audit, a single file copied to its final storage is considered equivalent to an AIP. The ICGC has the technical definitions of all file formats used in production (4.2.2) and links metadata with data through the file name (4.2.1.1), which also serves as a unique identifier (4.2.4.1.1).

Currently the IT department preserves all digital objects that it receives from creators and does not refuse any, which is why criterion 4.2.3.1 has not been evaluated.

The different types of preservation metadata that are part of the preservation description information at the ICGC are dispersed. Therefore, only part of the process of obtaining this information is documented (4.2.6.1) and only part of the metadata obtained (4.2.6.2). ISO 16363 also requires preservation description information to be linked persistently to the content while at the ICGC only the descriptive metadata that would be part of the PDI is related to the data (4.2.6.3).

The ICGC did not determine a designated community, but the users of its website are general users, internal staff and commercial clients. The ICGC does not take the general users into account in its decision process about the file formats and distribution channels (4.2.7), but an involvement process exists for internal staff.

The IT department does not check if sent information is correct (4.2.8) (format conformance, presence of metadata…) nor is it able to check if the information is complete (complete logical series with all the accompanying material) due to the fact that the responsibility to control what is copied lies with the data creator (4.2.9).

4.3. Preservation planning

The major difficulty in auditing this section is the lack of a preservation plan (4.3.1), nevertheless some aspects that would be part of the preservation plan are managed and controlled by the ICGC through other channels.

Even though for some digital preservation risks, such as hardware obsolescence, there are processes to prevent them, there is no exhaustive identification of preservation risks (4.3.2).

Currently the IT department trusts data creators to alert it to format obsolescence (4.3.2.1); This policy is insufficient because in the medium or long term the archive will contain file formats that have to be rendered but that are not in use in production any more, and therefore criteria for production and use will not always be the same.

The ICGC implements mechanisms to control and adjust its actions that work for hardware migration but so far, these mechanisms are not used to control file format obsolescence (4.3.3.1). Because the ICGC does not implement mechanisms for format migration, it cannot evaluate the trustworthiness of such migration. Neither can the ICGC document it. Currently this is a minor problem because digital stocks are relatively new (4.3.4).

4.4 Archive information package (AIP) preservation

The ICGC does not use packaging or compression technology, which is why knowledge needed to open and render the AIPs is reduced to the technical specifications of the file format. All file formats are known in depth and documented by the ICGC (4.4.1.1). To improve the evaluation of this criterion the ICGC could assure that the file format specifications are stored together with the data they concern. Otherwise the ICGC is at risk of not being able to open files due to missing technical specifications.

The backup system checks integrity at the time of file copy. The file storage system can check integrity of data that is online, but not of offline data (4.4.1.2); this is why there is a risk for the ICGC to lose information due to file corruption.

Actions that can be performed upon the archived files are controlled by the access platform for the users (4.4.2.1).

4.5 Information management

The ICGC creates metadata based on the INSPIRE standard and adapted to its own requirements. Because there is no defined designated community, it is not possible to check if metadata is sufficient for them (4.5.1). The ICGC commits to answer the information needs of individual users when requested by them.

For dissemination purposes, metadata are associated to data in the form of a zip file (4.5.2). For internal purposes metadata or other descriptive information is associated to the data by the file name. The file name is a bidirectional link between description and data (4.5.3.1). Due to the crucial importance of the file name, it is essential that the ICGC carefully keeps all changes documented through time, which includes documentation of abbreviations, tables and symbols used in the file name.

4.6 Access management

Electronic data access is regulated. Producers interested in files stored at long term storage have to formally request access. This process is documented (4.6.2). It is not possible to overwrite an archived file. Physical access to the digital archive is only partly controlled and presents the risk of unauthorized people entering the area.

The ICGC subscribes to ISO 9000, the standard for quality management, when it comes to content distribution to external users. Incidents at external distribution go through registering and analysis, but, incidents in internal access and distribution are neither registered nor analysed. To guarantee the security of archives, the ICGC should document and update all processes related to incidents and standardize decisions to take in each case (4.6.2.1).

4.2.3 Section 5: Infrastructure and security risk management

5.1. Technical infrastructure risk management

The ICGC changes its technological infrastructure regularly. Nevertheless, there is no system implemented to warn about file format obsolescence in the archive (5.1.1.1.6). Lack of such a system can bring the ICGC into a situation where it holds file formats incompatible with the newer software it is using.

User groups are not yet part of the decision process about access technology (5.1.1.1.1) In their place, internal working groups evaluate technologies and formats, which is no guarantee of their recommendations being appropriate for future user needs, especially of external users. At the time of technology change in particular, it is important to involve the designated community to assure that it is capable to understand and render the content.

Economic funding of technological changes is not guaranteed, as required by criteria 5.1.1.1.4 and 5.1.1.1.8.

On the one hand, decisions over technological changes are based primarily on economic criteria and do not appropriately include parameters related to preservation risks such as file format obsolescence (5.1.1.1.7). On the other hand, changes to and updates of the security systems are based on a risk-benefit assessment (5.1.1.4).

There is a lack in documenting change in access processes, management and data security (5.1.1.6.1). Without proper identification of critical preservation processes and documentation of its changes, the ICGC will not be able to monitor and evaluate the effects it has on the accomplishment of preservation objectives.

Backup copies are synchronized and stored correctly at different geographic locations (5.1.2.1). Moreover, the ICGC can identify the last file copied in case of back-up interruption.

5.2 Security risk management

In case of major incidents, the ICGC risks losing more content than necessary due to lack of periodic evaluation of security risks such as environmental, political or financial factors, or due to a lack of emergency plans (5.2.1). Security management instruments such as emergency plans or documentation of emergency recovery proceedings should be stored next to the backup copies. In this documentation, the ICGC should define the roles of the staff members in case of emergency (5.2.3).

The ICGC has to make sure that knowledge of its staff is documented and available together with the long-term archival holdings and backup copies (5.2.4).

4.2.4 Conclusion of the digital preservation audit

The audit detected different problems. Generally speaking, there is the absence of a defined designated community for the long-term archival holdings and the lack of a preservation plan.

When it comes to organizational infrastructure, the ICGC is appropriately organized and stable, which is a good base for taking responsibility for digital objects in the long term. Nevertheless, consciousness of digital preservation risks could be better recognized in emergency planning, including extreme scenarios such as total shut down of operations or an earthquake.

With respect to digital object management, many processes at the ICGC could serve preservation processes, but are currently not applied to long-term archival holdings in a consequential manner. For example, map and metadata bundling is used currently for dissemination but not for archiving. Other processes, such as migration planning, are missing completely. Some responsibilities are blurred between staff of production units and IT departments. For example, it is not clear who should define the metadata to preserve together with each object or who should check for completeness of the files needed for interaction with the stored spatial data.

Regarding the risk management infrastructure, not knowing the designated community that the future archival holdings should serve handicaps the evaluation of some risk criteria. The ICGC is constantly bettering the documentation of processes, their changes and the internal knowledge base; nevertheless, these efforts should be widened to new areas serving digital preservation.

As a result of the evaluation, the audit team recommended a series of actions for the ICGC, of which we would like to highlight the definition of the designated community. The ICGC is hindered in implementing several other preservation actions spanning all three areas evaluated by the ISO 16363 norm because of not having its designated community defined.

5 Results and discussions

5.1 Users of geographic information

This section gives an overview of the user types and profiles. First, we explain how the characteristics that we call modules of the user profiles emerged. Modules are theoretical constructs that can be applied on user types and clusters. Second, the user types and clusters as predicted by the Delphi experts are presented. Four user clusters are chosen and analysed in detail by comparing them to the current reality of Catalonian legacy geodata users. Finally, we list available legacy geodata available to users in Catalonia that we should have in mind when analysing the data.

5.1.1 Modules of the user profiles

The questions for the Delphi study were determined by the properties we expected to isolate for each user group; the properties should allow for making decisions about appraisal of geodata. Therefore, properties were extracted from existing appraisal guidelines and best practices. The characteristics of the users were synthesized and named modules, following the concept of Giaretta (Giaretta, 2011). One or several characteristics can inform a module value. The sum of the module values is the profile of the user type or user cluster.

As follows, we shall explain how each module appeared. A first reference point was the thinning criteria of the Ellipse project. These include: duplication, redundancy, scale and acquisitions interval, as explained in chapter 3.2.2. In the context of the Ellipse project these criteria were intended to allow automatic elimination of data sets prior to appraisal. Ellipse found that, because none of them could really be applied objectively, thinning was discarded (Bos et al., 2010). Nevertheless, the criteria play a role in the individual assessment of data sets, which is why we use these criteria for the definition of modules of the user profiles. The time range module was also inspired by Ellipse. The Swiss project analysed the three stages of archival material: active, semi-active and definitive, and pointed out the importance for some data to stay in the semi-active stage for several years. The time range module should assist in determining how long data should stay in the semi-active stage and whether users would access data at any or either of the stages.

A second reference point was the properties the DANS data repository recommends taking into account when adapting selection to the designated community. These include: data presentation, format, and processing stage (raw or processed). Under 'presentation' we understand the rendered versions of geodata. The amount of context and additional documentation needed also plays a role (Tjalsma & Rombouts, 2011). A series of questions about the importance of the user group was added, following the example of Kärberg (Kärberg, 2014). The module 'interaction type' came from the interpretation of Elzakker's thesis about the users of maps and should support the modules of file formats and type of product (Elzakker, 2004).

By summarizing the inputs from the Ellipse project, DANS recommendations, Elzakker's thesis and Kärberg's approach, and by having the ICGC production chain in mind, we defined the following modules:

Module name	**Interaction type**
Definition	The interaction type explains how the user groups reuse data, what they do with it and with what purpose.
Origin	Elzakker, 2004.
Description and purpose	The interaction type should complement and confirm information about the products users choose and the file formats that would serve them. For example, if users look at represented geodata as maps online and, apart from zooming, do not interact with the information, we can conclude that raster images would serve the purpose. On the contrary, when users create vector products or query geodata they would certainly prefer intelligent data formats.
Delphi questions	Round 1 and 2, group 1, question 3.

Table 4: Describes the module interaction type, its definition, purpose and origin.

Module name	**Time range**
Definition	Defines the age range of data the user group is interested in.
Origin	Ellipse concept of enduring availability. [35]
Description and purpose	The time range helps to determine after how many years archived data should pass from current to semi-active and from semi-active to definitive archives.

[35] Translated from the German concept of *Nachhaltige Verfügbarkeit*.

Delphi questions	Round 1, group 1, question 10 and round 2, group 1, questions 10, 11, 12 and 13.

Table 5: Describes the module Time range, its definition, purpose and origin.

Module name	**GIS knowledge**
Definition	Indicates if and to what degree the user has technological knowledge to manipulate legacy geodata.
Origin	Giaretta, 2011.
Description and purpose	This supports the module about file formats, informs representation information for an OAIS and can assist the definition of service levels in an archive.
Delphi questions	Round 1, group 1, question 22 and round 2, group 1, question 29.

Table 6: Describes the module GIS knowledge, its definition, purpose and origin.

Module name	**Scales**
Definition	Determines the scales the user group is most interested in.
Origin	Ellipse thinning criteria ‘redundancy’ and ‘scale’.
Description and purpose	Redundancy exists when data is available both in an individual data set and included in another. In the case of maps, smaller scale data is included in larger scales when the smaller scale is generalized. Therefore, certain scales can be considered redundant and we will investigate which of the scales the users are interested in.
Delphi questions	Round 2, group 2, question 3 and again interviewing the relevant user groups.

Table 7: Describes the module scales, its definition, purpose and origin.

Module name	**Type of product**
Definition	Describes basic characteristics of data objects such as if they are organized in database form or individual files, or if data are in raster or vector form.
Origin	DANS guidelines ‘presentations’ and ‘processing stage’ and the ICGC’s production chain.
Description and purpose	User preferences can help determine whether a data object of the production chain can be discarded. We asked if users primarily need the raw, intermediate or end products; if they prefer the raw aerial photographs or the adjusted orthophotos; and if they would prefer querying vector databases compared to downloading vector layers and raster maps. The type of product also assists in determining appropriate file formats.
Delphi questions	Round 2, group 1, question 26. The Delphi experts were asked to consider whether use of geodata would shift from end product to increased use of intermediate products in the next 10 years (Round 2, group 1, question 25).

Table 8: Describes the module Type of product, its definition, purpose and origin.

Module name	**Acquisition interval**
Definition	Indicates the maximum number of years between legacy geodata versions that should

	be available for a user group to work on their projects. It concerns the snapshot frequency of databases or the retention interval of aerial photographs.
Origin	Ellipse thinning criteria 'acquisition interval'.
Description and purpose	The acquisition interval is an important instrument for archives to achieve smaller collections.
Delphi questions	Round 1, group 1, question 32 and round 2, group 1, question 39.

Table 9: Describes the module acquisition interval, its definition, purpose and origin.

Module name	**Significance of the user group**
Definition	Gives indicators about the significance of the user group, such as frequency of access to legacy geodata, data amounts requested, percentage of the whole potential user access and the group's impact on the economy.
Origin	Kärberg, 2014.
Description and purpose	The significance of the user group is relevant for the decision on the designated community because it influences the level of support that will be demanded and therefore the feasibility of preservation. The economic impact of the user group can also be an argument for choosing it as a designated community.
Delphi questions	Round 2, group 1, questions 6, 7, 16 and 32 to 35.

Table 10: Describes the module significance of the user group, its definition, purpose and origin.

Module name	**File formats**
Definition	Defines which file formats are used or would be helpful in problem solving for the user group.
Origin	Ellipse thinning criteria 'duplication' and DANS guidelines.
Description and purpose	Duplication occurs when a data set is present several times, either in analogue and digital form or as different file formats. Analysing the way people search, query and reuse data sets may give us an indication of which version and file format is most useful to them. Current practice of the ICGC is to offer data in various file formats. The goal is to limit the file formats for each data set.
Delphi questions	Round 2, group 2, questions 5, 6, 7, 10, 13 and 16. This overview of the market share of file formats should give us insight on what is available to users. Further on we asked the current users in the interviews which file formats they prefer.

Table 11: Describes the module file format, its definition, purpose and origin.

Information about the representations people use, the process stages they know and use and the context information and additional documentation they require for their projects was extracted from the interview and focus groups.

5.1.2 User types as predicted by the Delphi experts

The results derived from the answers of both Delphi groups are exposed as follows.[36] We shall first present the user groups and their characteristics in the order of the survey questions. In chapter 5.1.2.1 we will see how user groups were clustered. Eight experts named user profiles in response to a Delphi study pre-question. The answers provided the following suggestions for user groups, supposed to be interested in legacy geodata:

- General public
- Archaeologists
- Historians
- Geographers
- Lawyers
- Policy makers
- Information officers in culture and arts (i.e., filmmakers and museum staff)
- Emergency response planning members
- Conservation agents of the non-natural environment
- Environmentalists (public and private)
- Architects
- Institutions that update and maintain geodata
- Undergraduate teachers and students

Experts were asked to add to the list of user groups in the first round. The following user groups accrued from the first round:

- Geophysicists (including seismologists, geologists)
- Social scientists (including economists, journalists, statisticians, people involved in development)
- Commercial users (including banks, utility companies, retailers etc.)

These user groups were included in the second round. All user-related questions of the first round were repeated for these user groups in the second round.

For the following tables, if a question did not obtain four valid answers the table shows ‘Not enough valid answers’ or a question mark instead of an answer.

[36] Part of these results were the subject of a communication at the Symposium of Service Oriented Mapping in 2014, in Vienna. Locher, A.E. Characterizing potential user groups for versioned geodata. Presented at the Symposium of Service Oriented Mapping, Vienna, 26-28 November 2014.

The first question explored the main tasks the user groups do with the legacy geodata they find. Do they consult or visualize it or do they query and amend it? The following table shows the opinion of the first Delphi group after the second round. It is indicated as 'both' when there was no majority for either option:

User group	task
Social scientists	both
Geophysicists	query
Commercial users	query
General public	consult
Archaeologists	query
Historians	consult
Geographers	query
Lawyers	consult
Policy makers (government)	consult
Information officers in culture and arts (i.e., filmmakers and museum staff)	consult
Emergency response planning members	consult
Conservation agents (non-natural environment. i.e., ancient building inspectors)	consult
Environmentalists of natural environment (public and private)	consult
Architects and engineers	both
Institutions that update and maintain geodata	query
Undergraduate teachers and students	consult

Table 12: What do these user groups mainly do with geodata?

Most user groups are said to consult the data. This would allow for offering them geographic information in the form of raster images that are independent of complex databases or software. As expected, geophysicists, geographers and institutions that update and maintain geodata need to query the data. Commercial users need to connect their data to the geographic information or offer new services by analysing and reprocessing it. It may be surprising that the experts include archaeologists in the category of users that query the data, because we expect archaeologists to use legacy photographs to detect archaeological sites and mostly amend recent data. This highlights one of the problems of the Delphi process: the answers sometimes seem to suggest that the experts forgot that the study concentrates on legacy data and seemed to consider geodata in general.

For the second question, we asked the experts to estimate the percentage of access of each user group; they were told that the total should equal 100%. The following results come from Delphi group one and represent the median of all estimates. A question mark is shown for user groups lacking enough valid responses:

User group	% of access
Social scientists	?
Geophysicists	10.0%
Commercial users	15.0%
General public	23.0%
Archaeologists	5.0%
Historians	8.5%
Geographers	8.0%
Lawyers	?
Policy makers (government)	4.0%
Information officers in culture and arts (i.e., filmmakers and museum staff)	4.5%
Emergency response planning members	5.0%
Conservation agents (non-natural environments)	?
Environmentalists of natural environment (public and private)	2.0%
Architects and engineers	3.0%
Institutions that update and maintain geodata	10.0%
Undergraduate teachers and students	3.5%

Table 13: Percentage of access to legacy geodata of each user group.

The sum of the estimates for all user groups is more than 100%. Given that access would include also the percentage of the three user groups for which a valid result was not reached, the overall estimate must be considered too high.

For the third question, we wanted to find out in which map-age range the user groups are most interested. The following table shows the minimum and maximum age in years the first Delphi group estimated for the user groups. Where '999' is indicated, it means that this user group is interested in data as old as possible.

User group	Min. age	Max. age
Social scientists	?	999
Geophysicists	1	70
Commercial users	?	40
General public	1	509
Archaeologists	1	999
Historians	7.5	999
Geographers	1	150
Lawyers	1	50
Policy makers (government)	1	65
Information officers in culture and arts	1	750
Emergency response planning members	1	25
Conservation agents (non-natural environment. i.e., ancient building inspectors)	?	170

User group	Min. age	Max. age
Environmentalists of natural environment (public and private)	1	170
Architects and engineers	1	50
Institutions that update and maintain geodata	1	200
Undergraduate teachers and students	1	200

Table 14: Minimum and maximum age of data the user group is interested in expressed in years.

Some user groups, such as commercial users, lawyers, policy makers, emergency response planning members and geo-related architects and engineers are interested in rather young legacy data of up to 65 years. It is expected that they might never access this data through a long-term archive, but instead from data producers or services that also deliver the most current updates. It is important to add that the experts mainly agree that the availability of data influences the data age range in which a user group is interested. This means that some user groups would reach back even further, if data were available to them. These answers indicate that most users compare current with legacy data and do not only work with historic data, so that a service that provides both in the same place or through the same application would be appreciated.

Subsequently the experts were asked to select the user groups that might request the largest amounts of data (in bytes not in land coverage). The four most mentioned user groups were the following. Next to the name we state how many times that user group was selected.

- Geographers (10)
- General public (9)
- Social scientists (8)
- Institutions that update and maintain geodata (7)

We can contrast this information with the results of a question asked to the second Delphi group, where experts classified the user groups into the following four profiles:

Class	User group
frequent user, needs large amounts of data	Geophysicists Geographers Policy makers Institutions that update and maintain geodata

Class	User group
frequent user, needs small amounts of data	Social scientists Archaeologists Historians Environmentalists
infrequent user, needs large amounts of data	No user groups in this category
infrequent user, needs small amounts of data	General public Information agents in culture and arts Undergraduate teachers and students

Table 15 User groups classified by frequency of use and amount of data they need

For some user groups no classification dominated, but it could be determined that conservation agents of the non-natural environment probably were frequent users and emergency response planners probably were infrequent users. For commercial users, the expert opinions diverged and for lawyers we did not receive enough answers. It is striking that in one expert group the general public was considered to require large amounts of data, while the other Delphi group considered them to ask for small amounts. It might be explained by one Delphi group thinking of the general public as a whole, where all the data added up by the individual users makes up a large amount, while the other Delphi group thought of individual members of the general public, who probably will not ask for large amounts of data.

The next question was about the most required data set by the user groups. The following table shows the most and second-most mentioned data sets for the user groups:

User group	1st choice	2nd choice
Social scientists	?	?
Geophysicists	DB	VM
Commercial users	RM	VM
General public	RM	OR
Archaeologists	?	?
Historians	RM	DB
Geographers	?	?
Lawyers	?	?
Policy makers (government)	RM	?
Information officers in culture and arts (i.e., filmmakers and museum staff)	?	?
Emergency response planning members	?	?
Conservation agents (non-natural environments)	?	?

User group	1st choice	2nd choice
Environmentalists of natural environment (public and private)	?	?
Architects and engineers	DB	?
Institutions that update and maintain geodata	DB	RP
Undergraduate teachers and students	RM	OR

Table 16: Most used data sets by user group

DB = Vector database
VM = Vector map (shareable file)
RM = Raster map
OR = Ortho-corrected photographs
RP = Raw aerial photographs

The options 'lidar data' and 'digital terrain model' did not reach the first two positions for any user group. We believe this is partly due to the relatively recent appearance of lidar data and digital terrain models compared to other types of geodata. In this table, on the one hand, maps dominate over photographs. On the other hand, counting vector maps and vector databases together, vector data slightly dominate over raster maps. This question should support the answers about what type of file formats are acceptable for the user group.

We then asked about the percentage of users with knowledge of GIS within each user group. The estimates of the first Delphi group are shown in the following table and expressed as the median of all estimates:

User Group	% with GIS knowledge
Social scientists	17.5%
Geophysicists	?
Commercial users	12.5%
General public	5.0%
Archaeologists	20.0%
Historians	5.0%
Geographers	90.0%
Lawyers	?
Policy makers (government)	7.5%
Information officers in culture and arts (i.e., filmmakers and museum staff)	10.0%
Emergency response planning members	30.0%
Conservation agents (non-natural environment)	?
Environmentalists of natural environment (public and private)	30.0%
Architects and engineers	30.0%
Institutions that update and maintain geodata	95.0%
Undergraduate teachers and students	7.5%

Table 17: Percentage of users with GIS knowledge in each user group

As in Table 13, we believe these results to be optimistic. It is true that nowadays during academic training geographers will learn to work with GIS, but not all geographers of previous generations employ GIS at their workplace. In the architects and engineers group, there are probably more members that know CAD than GIS. Here the fact that we did not further define 'knowledge' limits our capacity to interpret results. Therefore, although the proportions between user groups are interesting, their explicit value is questionable. This question gives insight into the definition of the appropriate file formats and needed significant properties. This characteristic indicates also appropriate levels of service.

Subsequently, participants established a ranking of the user groups by how the economy would be affected if no legacy geodata were available to them. The user groups that would have a bigger negative impact on the economy if they had no access to legacy data were ranked at the top.

1. Institutions that update and maintain geodata
2. Geographers
3. Environmentalists of natural environment (public and private)
4. Architects and engineers
5. Geophysicists
6. General public
7. Commercial users
8. Policy makers (government)
9. Emergency response planning members
10. Social scientists
11. Historians and at the same level undergraduate teachers and students
12. Information officers in culture and arts
13. Conservation agents (non-natural environment)
14. Lawyers
15. Archaeologists

Geographers are a large user group that would have a big effect on the economy if they lacked legacy data. The impact of environmentalists probably lies more in the fact that catastrophes could not be predicted and prevented without time series. The same interpretation is made for emergency response planning members, but because this user group is significantly smaller and the consequences of not being able to predict are more local than in the case of environmentalists, the group has been ranked further down. The big impact of architects and engineers is related to the cost and scale of their work, because construction usually affects a wide community around the construction site. We think this is also true for geophysicists. Surprisingly, lawyers are ranked quite low considering that court decisions about property and construction can be quite expensive. This indicates a small user group with infrequent needs for small amounts of legacy data.

We also addressed the time interval of archived geodata the user groups might need. In the first round, we asked about geodata in general, and in the second round we asked about remote sensed data and vector data separately. Unfortunately, not enough experts were aware of this separation and did not answer the question about vector data in the second round. The estimates were made in years. The experience the Delph experts have with the various user groups has not been taken into account for weighting, because it would have made it impossible to obtain results for half of the user groups and the question about time interval preferences is crucial for archives. The results obtained in the second round are the following and represent the median:

User Group	Year interval
Social scientists	5
Geophysicists	5
Commercial users	0
General public	5
Archaeologists	25
Historians	10
Geographers	5
Lawyers	5
Policy makers (government)	10
Information officers in culture and arts (i.e., filmmakers and museum staff)	10
Emergency response planning members	0
Conservation agents (non-natural environment. i.e., ancient building inspectors)	10
Environmentalists of natural environment (public and private)	10
Architects and engineers	5
Institutions that update and maintain geodata	1
Undergraduate teachers and students	10

Table 18: Estimated maximum acquisition interval of geodata useful for each user group measured in years

The figure zero was used to express ‘This user group needs all updates’. The experts who estimated for both vector and raster data indicated in general bigger intervals for raster data or equal intervals compared to vector data. One expert thinks that social scientists, the general public, archaeologists and lawyers need more frequent updates of raster data than vector data. The other expert thinks this is true only for geophysicists.

We did not define ‘raster data’ in the beginning of the survey form. It was intended to mean digital geographical information served in raster file formats. Analysis of the open questions revealed that some experts understood raster as data captured in raster formats such as lidar and other sensor data but excluding photography or maps rendered as raster. This ambiguity should be taken into account when interpreting the results of this question.

From the first round, we also got insight into how user groups might be affected by the elimination of smaller scale data even though all versions of larger scales are preserved. None of the user groups was considered to be unaffected. User groups that are slightly or severely affected are grouped as follows:

Slightly affected user groups	Severely affected user groups
General public, Information officers in culture and arts (i.e., filmmakers and museum staff), Environmentalists of natural environment (public and private), Undergraduate teachers and students	Archaeologists, Historians, Geographers, Lawyers, Policy makers (government), Emergency response planning members, Conservation agents (non-natural environment. i.e., ancient building inspectors), Architects and engineers, Institutions that update and maintain geodata

Table 19: Assorting the user groups by how much they would be affected if smaller scale data was not available.

5.1.2.1 Similarity of user groups

Finally, Delphi group two was asked to match user groups that might have similar necessities about functionalities in legacy geodata. Every user group could be matched to any other with the most similarities, so that an expert could set up a maximum of two associations for a matching pair. Each link between two user groups augmented the edge weight between them. Similarity is a non-directional relationship, because if A is similar to B, B is also similar to A. Because in our case all links are non-directional, integer and positive, the weighted graph can be mapped easily to a non-weighted multigraph (Newman, 2004). Therefore it is possible to apply the Louvain method for community detection, which is based on modularity optimization (Blondel et al., 2008). Modularity is the degree of connectivity between the nodes of a cluster compared to the nodes of other clusters or communities. Modularity is high when nodes in a same cluster are highly connected while links to other clusters are sparse. We expected to find communities in the network and applied the Louvain clustering method, which is available in the Gephy software package. The method detected six clusters in the network diagram, which were then partitioned by modularity class and expressed in different colours (Figure 8). For better visualization of the graph, edges with weights of one point are masked.

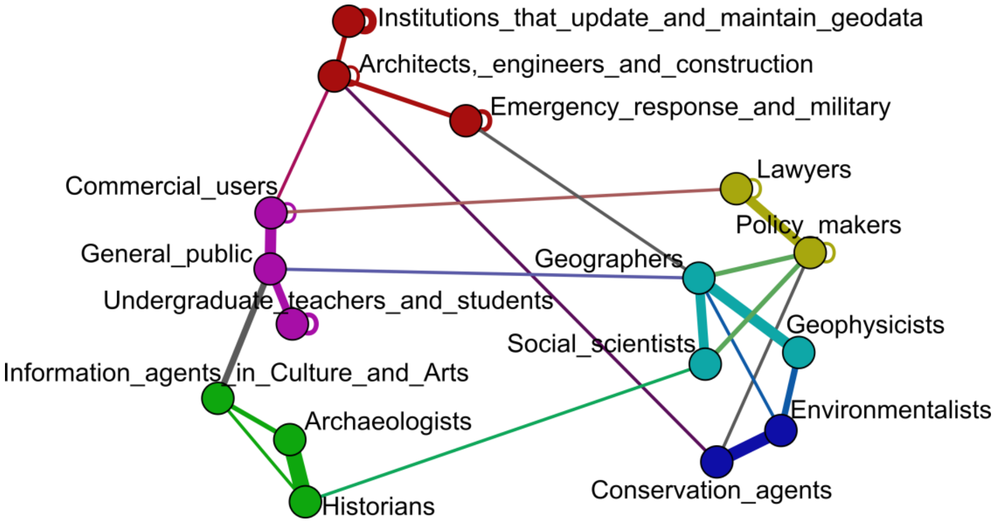
Institutions_that_update_and_maintain_geodata
Architects,_engineers_and_construction
Emergency_response_and_military
Lawyers
Commercial_users
Policy_makers
General_public
Geographers
Undergraduate_teachers_and_students
Geophysicists
Social_scientists
Information_agents_in_Culture_and_Arts
Environmentalists
Archaeologists
Conservation_agents
Historians

Figure 8: Network where the nodes are the user groups and the edges represent similarity in needs for functionality in a data archive for legacy geodata. Thickness of the edges expresses the degree of similarity and colours, the affiliation to a cluster

As shown in Figure 8, there are several user groups with similar expectations towards legacy data sets. The clusters – also called communities – which were formed by the second Delphi group, are

- Archaeologists, historians and information officers in culture and arts
- Policy makers and lawyers,
- General public, commercial users and undergraduate teachers and students
- Geographers, social scientists and geophysicists
- Environmentalists of the natural environment and conservation agents of the non-natural environment

The cluster marked in red has weak edges between its members, which mean that those user groups have specific needs not so similar to each other. The red cluster is formed by

- Institutions that create and maintain geodata
- Architects, engineers and construction agents
- Emergency response planning teams and military

The social scientists, geographers, geophysicists, environmentalists and conservation agents are all connected through similarity, but the group has no centre node like the cluster around the general public. It has a linear form in which each node is linked to its previous and following node, so that we can assume that the user groups at the extremes do not have much similarity. This is confirmed by the modularity information of the nodes expressed by colours, which shows two clusters: on one extreme the geographers, geophysicists and social scientists and on the other extreme the environmentalists and conservation agents of the non-natural environment.

The case of the geographers, tied to both the social scientists and the geophysicists, shows the multidisciplinary nature of this subject. A geographer can specialize in human geography as well as physical geography. The manner in which the experts expressed the geographers' similarity with one or another related user group was affected depending on which specialists they had in mind.

The similarity of three of the groups – institutions that create and update geodata, architects and emergency response planning teams and military – is low, so the edges between them are very thin (Figure 8). However, they have even lower modularity to other user groups which is visualized by them being partitioned in a separate cluster.

5.1.2.2 Selection of user clusters for profiling

Hereafter we will discuss the characteristics of the user groups in the same cluster, to see if they should be considered for the focus group interviews.

We can see that the general public is considered similar to a variety of other user groups. In particular, commercial users and undergraduate teachers and students are in the same cluster. Together with its cluster members it would account for 41.5% of the access to legacy geodata, which makes this group an interesting target in terms of brand awareness of a potential service provider. For its importance and interest in legacy data, all user types in this cluster are retained as a user profile under the name 'general public'. Historians and archaeologists are very much seen as needing similar functionalities in legacy geodata. They are also considered similar to information officers acting in culture and arts. All three need ranges of recent to very old data and would be able to work with larger time intervals of 10 to 25 years. Another similarity is that they would all have a low impact on the economy if no legacy geodata were available, which might be a disincentive for a commercial service provider, due to related lack of economic power/pressure by the group. Nevertheless, for their pronounced interest in legacy and historic data they will be interviewed and referred to as the user profile of ‘historians’. The policy makers and lawyers user groups are seen to have similar needs for functionalities. Although we do not have much information about lawyers due to the fact that few experts have experience with this user group, the related policy makers are estimated to be frequent users and to have a big negative impact if no legacy geodata were available. These two characteristics and the fact that they represent political and economic power mean that they can make an interesting market segment and are important to interview. Policy makers and lawyers will be grouped under the user profile name ‘urban planners’. Social scientists, geographers and geophysicists are thought to be frequent users who would need large amounts of legacy data. Both characteristics make them an interesting target for a service provider and worthy of interviewing in the focus groups. Even though we have no reliable information about the most desired data set by social scientists and geographers, we know that geophysicists would prefer to work with a geographic database or vector data. That this applies also to the other two user groups can be affirmed by the fact that all of them are assumed to query legacy data, which is easier with vector data. In the case of choosing this cluster as a designated community or market segment, on one hand, service providers could serve vector data, which is more complex to archive. On the other hand, they could count on a higher knowledge of GIS systems by their clients, falling between 17% and 90%. This cluster accounts for more than 18% of the access to legacy data. We will refer to this cluster as the user profile of ‘geographers’. Analysing the characteristics of environmentalists and conservation agents of the non-natural environment, we see that the needs of these user groups would be satisfied with a service designed for the general public. For example, all three groups are said to mainly consult data and the acquisition interval demanded by the general public is shorter than the one predicted for environmentalists and conservation agents (see Table 18). Therefore, we will not further analyse user types of this cluster. If institutions that create and maintain geodata do not hold the intellectual property of the legacy geodata they consume, or if they do not want to archive it by themselves for the period corresponding to their interest (200 years), they would be an interesting partner for a long-term service provider due to their economic impact and need for frequency and access. Nevertheless, we excluded such institutions from focus group interviews, because the ICGC belongs to this user group, and it is much more sensitive to its own needs through its internal processes than could be detected by a focus group. Finally, architects, engineers and construction workers, and emergency response planning teams and military are interested in geodata up to 25-50 years old and require rather small amounts of data. Therefore, we have to consider that they might never use the actual long-term data archive but get this data directly from the producer or through other short-term geographical service providers. Therefore, members of the red cluster (see Figure 8) are not taken into account for focus group interviews.

5.1.3 Legacy geospatial data available in Catalonia

Several sources of legacy geodata are available to current users in the area of Catalonia. The Map Library of Catalonia (*Cartoteca de Catalunya*) has a huge amount of historic and antique maps on paper, but also all paper editions of recent years. Many of these maps from the 16th century up to recent years are scanned and made available through the website of the institution. Maps that are not of their own production and that are younger than 70 years are not online for copyright reasons. Aerial photographs are also available on demand at the ICGC, and since 2012 all orthorectified images of historic and current flights undertaken by the ICGC are online and ready to download. The earliest aerial photographs available online are the American flight – series A of 1945. Until 2015 the oldest digitized flight was the American flight – series B of 1956.

The graphic department (*Unitat Gràfica*) of the National Library of Catalonia also has an important collection of historic and recent maps, as it acts as legal depository for edited maps. Some of its older maps from the 17th century up to the middle of the 20th century are available online.

The public administration of the metropolitan area of Barcelona (*Àrea Metropolitana de Barcelona*) offers historic aerial photographs of ten flights carried out between 1956 and 1992 and five corresponding orthorectified image series on its website, with the possibility to consult online and download. There are also aerial photographs produced in 1954 by the tax office at a higher resolution than the American flight, but they are neither digital nor georeferenced.

The Spanish National Geographic Institute also offers online viewing and downloading of historic aerial photographs from 1929 on, scanned cartography starting with handmade drafts from 1870 and the digital edition of the Spanish National Atlas from 1955 to 2008.

Legacy cadastral information for Spain back to 2002 is available online.[37] Regional tax offices or regional archives offer older cadastral data as paper copies.

The library *Biblioteca Víctor Balaguer,* the library of the Catalonian hiking club (*Centre Excursionista de Catalunya*), the National Library of Catalonia and the ICGC's map library make part of their historic maps available online over the catalogue of the spatial data infrastructure of Catalonia. Metadata was adapted to the spatial data infrastructure's requirements and maps that were not yet georeferenced got a reference for that purpose (Montaner, Pascual, & Roset, 2014).[38]

The Government of Catalonia (*Generalitat de Catalunya*) offers a website with access and download options for thematic layers.[39] Most are current data with a few exceptions of layers merged from time series, such as the areas affected by forest fires in the years 1986-2015 and sound indexes from 2012 to 2017 and legacy layers such as land use in 1993.

Satellite images are available at Google Earth. Some regions have been covered several times and show images of earlier states of the surface, for some places as far back as 60 years ago. Some of these images at higher resolution are licenced from the ICGC.

Local authorities publish and archive large scale maps of their territory, which are available on demand. Most townships redirect demands for historic maps to the local archive where people can consult in situ. The few exceptions of municipalities that digitize their maps do so without an overall access strategy. An exception is the historic archive of Lleida, which offers its scans online in a catalogue for viewing and download. A few other private and public institutions have an agreement with the ICGC, which publishes their scanned historic maps on the ICGC website.[40]

Documents about land property in several archives can contain geographic information. Books scanned by general and map libraries might hold maps. Even though they might be available in online collections, if individual images of the book pages are not treated as maps and georeferenced they stay hard to find.

[37] http://www.catastrohistorico.com

[38] http://catalegidec.icgc.cat/SDIExplorer/cercaCataleg.jsp

[39] http://sig.gencat.cat/visors/hipermapa.html

[40] http://cartotecadigital.icc.cat/cdm/landingpage/collection/externs

5.2 Discussions of the main research goals

This chapter contains the answers to the two research questions. First, we concentrate on the resulting user profiles and their module values (research question one). Second, we synthesis the user input to amend appraisal guidelines (research question two). We end this chapter with a proposal to the ICGC for user-informed appraisal and selection of a part of their production.

5.2.1 Discussion of the user profiles

The profiles of the four chosen user clusters of legacy geodata are presented hereafter as they were indicated by the Delphi study and the focus groups. The module values of each profile realise the knowledge base we wanted to construct of legacy geodata users. To put to evidence where user input corresponds or diverges with the prediction of the Delphi, a summary of the resulting current user profiles is given in chapter 5.2.1.2.

5.2.1.1 Module values of the Delphi compared to those of the interviewed users

As explained in the methodology chapter, we examined four user clusters of the six that emerged from the similarity test in more detail. The four clusters comprise geographers, historians, urban planners and the general public. These generic names for the clusters are used, even though each covers all user types under its umbrella, as related in chapter 5.1.2.2.

The following tables show the names of the modules, the Delphi prediction and the user input. Where appropriate, we used some answers from the questionnaire to illustrate or support the cases made by the interviewed user groups. The goal of this exercise is to see if the Delphi profiles correspond with what we currently see in user behaviour and needs, as expressed by the representative users in the focus groups. Each of the four chosen user clusters is presented individually.

5.2.1.1.1 Geographers

This user cluster comprises geographers of human and physical geography, as well as social scientists.

Name of module[41]	Description as a conclusion of the Delphi study	User input about the module comes from groups and participants G1, G2, G6, G11 and G13.
Interaction type	The principal tendency in this group is to query and amend geodata. Abbreviated module value: Query	G1_P1 studies evolution of vegetation and land use and extracts information from aerial photographs to create maps. The participant teaches this subject as well. G1_P2 studies the change of river flow in Switzerland. Swiss cartography is chosen for its quality and abundancy and serves to reach conclusions for geological structures in other parts in the world. Legacy maps are used also to reinforce results from sediment analysis. G1_P3 does not need historic photographs for teaching. He uses old photographs to date geomorphological events especially erosion processes. G2_P2 studies soil use. G2_P1 makes his students use legacy geodata. G6_P2 consults historic aerial photographs to explain the current state of mountain landscape and has also used oblique photographs to translate elements to maps. G6_P3 consults and amends data. G6_P4 vectorises old maps to compare with newer maps.
Time range	Interest in legacy geodata ranges from most recent to 70 years back (geophysicists), 150 years back (geographers) and as old as possible in the case of social scientists. Abbreviated module value: Up to all data	G1_P2 goes back to the oldest maps available, from the 16th century. G1_P3 says that the older the photographs the better for his research. G1_P3 would like to go back further than the 1950s, but there is no documentation available. G2_P1 on the occasion of a celebration is working with maps from the 18th century. G6_P1 is interested in the last 300 years but would like to analyse place names from the beginning of humanity. G6_P2 values the photographs from prior to the American flight less because they do not cover the whole territory of Catalonia. G6_P3 works with the period of the 18th century up to current date. G6_P4 studies the entire range from 15,000 years ago up to our date. G11_P1 uses maps from the 18th century. G11_P1 and G11_P2 would prefer aerial photographs of the 18th century if it would exist, because it is more objective.

[41] For definitions of the modules, please see chapter 5.1.1

Name of module[41]	Description as a conclusion of the Delphi study	User input about the module comes from groups and participants G1, G2, G6, G11 and G13.
GIS knowledge	A large number of geographers have GIS knowledge (90%) while only 17.5% of social scientists are estimated to have GIS knowledge. Abbreviated module value: High to medium	G1_P1 and G1_P2 georeference old maps. G1_P3 does not resolve the technical parts of his investigation himself. G2_P2 creates cartography from the aerial photograph. G6_P3 introduces information from various sources into a GIS and edits new maps of historic situations. He changes data from one coordinate system into another if necessary. G6_P1 has never used GIS, but makes others use it. In addition to GIS, G6_P2 handles CAD. G6_P3 and G6_P4 have basic knowledge of CAD. G6_P3 works with MiraMon.
Scales	Geographers would be severely affected by the elimination of smaller scale maps, even though larger scales are preserved. Abbreviated module value: Needs all scales	G1_P1 wishes to work with 1:2,000 or 1:5,000. G1_P3 says that there could be need for smaller scales and if that scale does not exist, his team would recreate it based on the larger scale, but he is not happy with this extra work. G6_P1 and G6_P4 want to have all scales, to be able to zoom in and out. G6_P1 is most interested in 1:1,000 on up to a global scale. G6_P3 thinks that it is not necessary that all scales exist. He says that they would create information of all the scales about the topic they research. He also recognizes that especially for images, having a high resolution might be too much for the computer to handle. G11_P1 says he works with topological maps of 1:25,000 and 1:50,000 scale. He states that zooming in to more than 1:5,000 is quite exceptional for his geological studies. G11_P3 has used 1:500 and 1:1,000 scale maps. G11_P3 says that for geophysics sometimes differences of 1 or 2 meters in the Digital Terrain Model are significant.

Name of module[41]	Description as a conclusion of the Delphi study	User input about the module comes from groups and participants G1, G2, G6, G11 and G13.
Type of product	Geophysicists would prefer to use databases and vector maps. We do not have similar information on Geographers and Social scientists, but because they also query and amend data, they probably would also use databases. Abbreviated module value: Databases and vector data	G1_P1 uses aerial photographs most of all. The analogue original, if it exists, is always of better quality than the scan. G1_P3 uses oblique photographs. G1_P1 thinks that the raw photograph is not necessary any more when the orthophoto is made correctly and is accessible. G6_P1 needs thematic information that is not available on topographic maps such as type of cultivation, but wishes to have more place name information. G6_P3, in addition to aerial photographs, uses thematic maps and toponymy. The information that G6_P4 need is easier to detect in thematic maps than in aerial photographs. Some information of the basic topographic map can be useful such as place names and transportation layer, but the participant does not need it as vector data because he interprets it in combination with other textual information. G6_P3 talks about an example where his department saved the process and control points for creating orthophotos but could not save the orthophoto itself due to lack of storage space. G6 in general is opposed to elimination of originals. G11_P1 uses aerial photographs to study geological processes such as faults, folds and river flow. On more recent photographs, urbanization is covering the view of the geological structure. They use maps from the 18th and 19th century as well and verify their accuracy by taking samples from the ground.
Acquisition interval	The maximum (time between) acquisition intervals for Geographers, Geophysicists and Social scientists is estimated to be five years. Abbreviated module value: 5 years	G1_P1 thinks that the acquisition interval is the same as the frequency of update regulated by Inspire. G1_P2 says that if for financial reasons this frequency cannot be archived digitally at least a paper version should be preserved. G1_P1 says that at least the end of project (closed layer) should be saved. G1_P1 and G1_P2 finally agree on an interval of 5 years for topographic maps, but for satellite imagery the minimum is to have a copy every year. The twenty-year interval between the year 1956 and 1977 is to long for G2_P2, which is why he works on recovering an aerial photograph from the 1960s. G6_P4 wishes to have a thematic map of her subject every 10 years. G6_P3 asks for annual copies of the information for the most current 100 years. If he could have aerial photographs of every month of the year he would save it as well. G6_P1 says that the snapshot frequency of vector databases should adapt to the speed the topic evolves. G6_P2 adds that even for the same topic, the evolution speed can vary significantly in time or between regions and gives as example the development of residential areas in the Pyrenees.

Name of module[41]	Description as a conclusion of the Delphi study	User input about the module comes from groups and participants G1, G2, G6, G11 and G13.
Significance of the user group	Geographers and Geophysicists together make up 18% of all access to legacy geodata. Both groups demand large amounts of data and are frequent users. Both are in the upper range of user groups concerning their negative impact on economy, if legacy geodata were missing. Social scientists are in the middle range. Abbreviated module value: High	For G1_P1 all and for G1_P2 the majority of projects include historic information. The closer the documents are to the present time, the more they are consulted because they are easier to obtain. For most of his work G2_P2 needs recent cartography. G6_P1 works daily with legacy data. G2_P2 makes his students study the evolution of the population in a certain town. G6_P4 does not consult old maps, because for the period the participant is interested in they do not exist.
File formats	No user specific predictions about file formats were made by the Delphi experts.	G1_P1 suggests archiving image files in TIFF or GeoTIFF and vectoring information as Shapefile, because it is a de facto standard. Personally, he would prefer DXF. G1_P1 says that the maps are produced in DGN (a CAD format for Microstation) which corresponds to the original and insists that the original should be preserved. G1_P3 says that he does not know about file formats. G1_P2 thinks that Shapefile might be easiest to work with because all programs can read it. G6_P2 is not worried about the file format as long as it offers the best resolution possible. He comments that if it is vector data, it should stay in a vector format. G6_P3 asks for JPG because it generates smaller files than TIFF and his computer otherwise is too slow. The resolution of JPG is sufficient for him. G11_P1 downloads PDF because he does not use GIS. G11_P3 also works with PDF and if he has to draw on it uses Photoshop. G11_P2 uses Shape and thinks that PDF can be imported into MiraMon and ArcGIS.

Table 20: Characteristics of the user groups clustered around Geographers.

Additional comments of relevance: G11_P1 and G11_P2 state that they work sometimes on behalf of government bodies to resolve a specific problem and sometimes on research projects. The group G6 is worried about the quality of georeferences.

5.2.1.1.2 Historians

This user cluster comprises historians, archaeologists and information agents in culture and arts.

Name of module[42]	Description as a conclusion of the Delphi study	Input from the following participants: G1_P4, G5_P1, G5_P2, G5_P3 and G12. In addition, we add the statements of two respondents of the initial questionnaire who see their map use as historic.
Interaction type	Historians, information agents in culture and arts and archaeologists are said to consult and query geodata. Abbreviated module value: Consult and query	G1_P4 uses maps and photos on paper for teaching; he and his students compare documents of different ages by eye. He wants to see how geodata is represented. G5_P1 studies the context of map creation and distribution. G5_P2, with her background in architecture, vectorises architect plans. The participant amends the plans with information from real world observation and other sources and publishes for museums and expositions. She is aware of the possibilities of data querying in a GIS and has ideas for research that would need querying, but her data is not vectorised. Participant G5_P3 produces written documents (articles and a book). In the course of his research an architect in his working group drew new architectonic plans and maps of houses in the core township. G12 downloads, compares and consults legacy geodata to identify possible cultural goods and archaeological sites but rarely loads data into a GIS.
Time range	Interest in legacy geodata ranges from most recent to as far back in time as possible. Abbreviated module value: Up to all data	G1_P4 is interested in the context of map creation of any age and century. G12 consults any maps, independently of the age. G5_P2 specializes in the 5th and 10th century but needs documentation of intermediate ages as well, to detect evolution in construction. G5_P3 does studies that go back to as far as the Middle Ages, but is also involved in recent history. S2 consults maps older than 100 years and S9 compares maps that are from 10 to more than 100 years old.

[42] For definitions of the modules, please see chapter 5.1.1

Name of module[42]	Description as a conclusion of the Delphi study	Input from the following participants: G1_P4, G5_P1, G5_P2, G5_P3 and G12. In addition, we add the statements of two respondents of the initial questionnaire who see their map use as historic.
GIS knowledge	It is estimated that 5% of Historians, 10% of Information agents in culture and arts and 20% of Archaeologists have GIS knowledge. Abbreviated module value: Medium to low	G1_P4 mentions analogue comparison of aerial photographs. G5_P1 has advanced knowledge in technology due to his background in geographic engineering and his experience in metadata crosswalks. G5_P2 has a background as architect and knows CAD. G5_P3 says that the technical parts of his studies were accomplished by others. S9 will create new map representations of the information found in legacy geodata.
Scales	Both Historians and Archaeologists would be severely affected by the elimination of smaller scale maps, even though larger scales are preserved. Abbreviated module value: Needs all scales	G5_P2 needs very large scales 1:100 and 1:200. G5_P3 studies urban town centres and the evolution of individual houses. G12 is using cadastre data that are also at a large scale. For them the elimination of derived scales is not a problem. G5_P1 would like to have excerpts of all resolutions to study their evolution. The scale is only somehow important to S2 and S9.
Type of product	Historians would prefer to use raster maps but also vector databases. We do not have similar information on archaeologists. Abbreviated module value: Raster maps and databases	G1_P4 mentions use of aerial photographs for teaching but is interested in any type of product as an object of study: topographic and thematic maps, plans and picture cards. Several participants use cadastre data. Partly because, at the resolution they need, there are no topographic maps available. G5_P3 also mentions aerial photographs and says that if old maps existed he would use them. All participants rely heavily on additional sources, such as the *Registro de la propiedad* [Public Registry of Real Estate Property], notarial documents, birth and family registries and non-aerial photographs. Toponymy is very important to S2.

Name of module[42]	Description as a conclusion of the Delphi study	Input from the following participants: G1_P4, G5_P1, G5_P2, G5_P3 and G12. In addition, we add the statements of two respondents of the initial questionnaire who see their map use as historic.
Acquisition interval	The maximum acquisition interval for historians is estimated to be 10 years and for archaeologists 25 years. To suit all users in this group, a 10-year interval is needed. Abbreviated module value: 10 years	G12 feels that snapshots should be taken more frequently than all 10 years. G5_P1 suggests different time intervals for every layer. The participant estimates that a 10-year interval for place names and a 1-year interval for streets would be enough. Additionally, snapshots of layers for which the first data capture has not finished yet should be taken in more regular intervals; this allows users to know at what time which information was available. G5_P2 feels that a snapshot interval of 2 years is as random as 5 years, as the participant needs evidence of every time a change is made and changes take place at irregular intervals. G5_P3 does not talk on that subject.
Significance of the user group	All three groups together make up 18% of all access to legacy geodata. All three demand small amounts of data; historians and archaeologists frequently and information agents in culture and arts unfrequently. Additionally, all three groups are in the lowest range of user groups concerning their negative impact on economy, if legacy geodata were missing. Abbreviated module value: Medium	Use of maps as information sources is minor compared to all the other sources historians need. This is even true for the historian that studies a specific old map and those who study the history of mapmaking. The former has his own copy of the map he accesses regularly and apart from this would mostly use documents that contextualize it. The latter accesses legacy geodata more often, but says that having samples would probably be enough for his work. All other reported research projects covered a very local area.
File format	No user specific predictions about file formats were made by the Delphi experts.	G5_P1 is very happy with TIFF but asks for more resolution than 300 dpi. When G1_P4 scans old maps, he uses JPG.

Table 21: Characteristics of the user groups clustered around Historians.

Additional comments of relevance: Participants in G12 produce scientific and documentary products and work also for government institutions.

5.2.1.1.3 Urban planners

This user cluster comprises the following user groups: Policy makers and urban planners and lawyers.

Name of module[43]	Description as a conclusion of the Delphi study	Input from participants in focus groups G3, G4, G8 and participant P2 of G2.
Interaction type	Policy makers and Lawyers are said to consult geodata. Abbreviated module value: Consult	The participant group G3 queries and analyses geodata. This is a transversal service they offer to local government departments. In support of a cultural and historic project they reconstructed and vectorised local authority border lines over the years to project them onto current maps. The study explained why certain planning decisions that might still be felt today were taken while the city expanded. G2_P2 creates and amends vector data and compares surfaces for statistical use. G4_P1 builds GIS projects on base maps from the ICGC. G8_P2 finds it very useful to be able to compare with the transparency tool. Participants in G8 ingest all flights from the ICGC into their information system. They query, amend and publish electronically.
Time range	Interest in legacy geodata ranges from most recent to an estimated 65 years back, (50 for Lawyers). Abbreviated module value: Up to 65 years	Members of G3 mapped their town at a state in Roman times. For a study on construction permits, G3_P3 had to compare data of the last 10 years. G4_P1 consults back to the year 1956 because previous aerial photographs do not exist. Ideally, she would go 100 years back, but says 50 would be enough. G8_P1 was engaged in a project to redraw the town at different points in time: the roman, medieval and industrial times and for 1960-1970. In another project, he often used a map from 1936 to localize country houses. G2_P2 uses the first aerial photographs that are available in his area of interest: 1956. He says that for urban planning, 50-60 years back is enough.

[43] For definitions of the modules, please see chapter 5.1.1

Name of module[43]	Description as a conclusion of the Delphi study	Input from participants in focus groups G3, G4, G8 and participant P2 of G2.
GIS knowledge	It is estimated that 7.5% of policy makers have GIS knowledge. No data is given for Lawyers. Abbreviated module value: Low	All participants of G3 have advanced technological knowledge of database creation and offer geographic services and information to third parties. G4_P1 does not manage geodata herself but coordinates a group that uses GIS and CAD. She has a background and experience as a cartographer. G8_P1 has put up the map server and helps people from other departments to upload their own data. G8_P2 and G8_P3 have studied geography and taken classes in GIS. G2_P2 creates maps out of aerial photographs and knows how to georeference and orthorectify them.
Scales	Both Lawyers and Policy makers would be severely affected by the elimination of smaller scale maps, even though larger scales are preserved. Abbreviated module value: Needs all scales	G3 is responsible for an area enclosing several townships and works with different scales. 1:500 is necessary for urban areas, while 1:2,000 is enough for forest. They are opposed to elimination of smaller scales basically because larger scales do not cover the whole territory. G4 works with all scales from 1:1,000 to 1:50,000, but the most important to them are 1:5,000 and 1:25,000. 1:1,000 is less important to them because this scale does not cover the whole territory. Territorial coverage is important. G2_P2 needs 1:1,000; other scales he can regenerate if necessary. But he would be happy to have a smaller scale additionally.
Type of product	Policy makers would prefer to use raster maps. We do not have similar information on Lawyers. Abbreviated module value: Raster maps	G3_P3 studied urban development through aerial photographs and old maps. G3_P2 verifies property information with cadastral data, old maps and aerial photographs – whatever is available. G3_P1 used additionally photographs of streets and buildings. G4 would be happy with the orthophoto and the vector map data. The participant considers them as final products. G8_P1 uses more maps for verification purposes of house numbers. G2_P2 uses all types of products: orthophotos, maps, topographic bases and thematic data.
Acquisition interval	The maximum acquisition interval for Lawyers is estimated to be 5 years and for Policy makers 10 years. To suit all users in this group a 5-year interval is needed. Abbreviated module value: 5 years	G3_P2 prefers to save data sets at the final point in production (at irregular intervals). The closing of a project would be the trigger of archiving it. In addition to the final project, his department permanently saves information every three months. G4_P1 and G8_P1 sometimes need geodata from a specific date in the past for legal purposes, therefore having a history of all changes would be useful to them. Participants in G8 make yearly copies of their own geographic production. They acquire all aerial photographs from the ICGC. G8_P1 is opposed to selection.

Name of module[43]	Description as a conclusion of the Delphi study	Input from participants in focus groups G3, G4, G8 and participant P2 of G2.
Significance of the user group	Policy makers are source of 4% of all access to legacy geodata. They are frequent users who demand large amounts of data. We have no information on the percentage of access and the frequency of use concerning Lawyers. Policy makers are in the middle and Lawyers in the lowest range of user groups concerning their negative impact on the economy, if legacy geodata were missing. Abbreviated module value: Medium	G3 are recurring users of historic cadastral data. G3_P2 says that the local authority departments as a whole use legacy geodata every day. The use by the analytical department depends more on the current project. G4 says that the frequency of access to legacy geodata is currently limited because people know that either they will not find the information, will not be able to open it or do not know where to look for it. Legacy geodata could be something the department consults every day, if it were easily available. G8_P2 uses legacy geodata with very low frequency (2 or 3 times a month) but thinks that the urban planning department of the town uses it much more. The participant is right now on a project where he uses historic data daily, but the project will last for about 1 or 2 months. G2_P2 mentions that legacy geodata is used regularly in his work.
File formats	No user specific predictions about file formats were made by the Delphi experts.	G4 supports any file format and transfers or joins any of them into the same GIS environment. If only one is available, they would choose Shapefile because they know how to transform it into their currently used formats. For raster data they prefer MrSID, because TIFF results in larger files and the detail that can provide MrSID is sufficient. G8 is OK with any file format as long as it can be incorporated into their Web-map system. Other departments serve data in Shapefile format or as Oracle elements. G2_P2 prefers GIS file formats such as Shapefile and geodatabase but also uses CAD data. For raster data, he uses Landsat, TIFF, MrSID and less frequently JPG. G2_P2 is restricted by the capacity of the receiving computer that sometimes depends on light formats such as MrSID. He is unable to cope with superseded or little known file formats.

Table 22: Characteristics of the user groups clustered around Policy makers and Lawyers.

Additional relevant comments: G3_P2 worked together with the department of archaeology of the University. They also studied the evolution of forest coverage compared to pastoral surface in a project related to fire prevention.

5.2.1.1.4 General public

This user cluster comprises members of the general public, commercial users and undergraduate teachers and students.

Name of module[44]	Description as a conclusion of the Delphi study	Input made up of interview group G9 and G10 and participants by email G7 and G14
Interaction type	General public and undergraduate teachers and students are said to consult geodata while commercial users query and amend. Abbreviated module value: Consult and query	G7 studies change in rural pathways through time. He views orthophotos and digitized old maps online. He also sometimes downloads and rarely prints out maps but never queries databases. G14 downloads and amends data and edits new maps. G9_P1 georeferences information, first on paper maps and later with the help of Google Earth. G9_P3 downloads data and consults online but also likes to go to physical archives. G9_P2 traces historic hiking paths on Google Earth with the help of textual descriptions from early 20th century and old maps. G9_P3 and G9_P1 acquire paper maps. G10_P1 consults maps online and on paper, in books. On screen, he wants to zoom in and have interactive maps to see the course of historic processes such as a battle. G10_P2 consults thematic layers on Google Maps in regions of current conflict or natural disaster. He rarely downloads things, maps even less. G10_P3 does not search historic information on the Internet, but in books. G10_P4 consults maps and photographs available in digital collections of libraries. She might occasionally download a picture and save it or send it to a friend.

[44] For definitions of the modules, please see chapter 5.1.1

Name of module[44]	Description as a conclusion of the Delphi study	Input made up of interview group G9 and G10 and participants by email G7 and G14
Time range	Interest in legacy geodata range from most recent to an estimated 40 years back for commercial users, 200 years for teachers and students and 500 years for the general public. Abbreviated module value: Up to 500 years back	G7_P1 consults everything that is available from the first orthophoto up to the most recent and old maps as long as they are available at the ICGC website. G14 is interested to obtain data as old as possible. G9_P1 consults a local map from his township from 1914. G9_P2 would like to work with aerial photographs older than the American flight, if they had the same quality. G9_P3 resorts to maps from the 19th century to extract street names, but regrets that he only found this type of map for Barcelona. He would like to have 19th century-maps of many other regions. G10_P3 likes to see how places in Catalonia known by him evolved in the last 100 to 200 years. G10_P1 and G10_P3 are interested in Roman times. G10_P1 also in medieval times. The work of G10_P4 covered the Second World War, most of all. Sometimes the same area at different moments in time was desired to show the effects of a specific event.
GIS knowledge	It is estimated that 5% of teachers and students, 7.5% of the general public and 12.5% of commercial users have GIS knowledge. Abbreviated module value: Low	G9_P2 has worked with ArcGIS to georeference maps. He uses online tools from the ICGC to compare and switch between maps and aerial photographs. G9_P1 and G9_P2 have websites where they publish findings. G9_P1 also uploads pictures to Panoramio but does not use a GIS. G9_P3 finds Google Maps easy to use, for moving around on the map and for finding details. G10_P1 and G10_P3 do not know what a GIS is. G10_P4 knows of a specific use case, such as localization of events on a map and G10_P2 has considered integrating a GIS in his knowledge management product, but neither uses a GIS.
Scales	Both General public and Undergraduate teachers and students would only be slightly affected by the elimination of smaller scale maps, when larger scales are preserved. Abbreviated module value: Can work with large scale maps	G7_P1 uses the 1:5,000 scale, but occasionally smaller scales for overview. He does not mind the elimination of derived scales but it is important to him that the preserved scales show homogeneous information. G14_P1 need a lot of detail. He is used to working with the 1:10,000 scale. G9_P1 prefers to work with a map scale of 1:10,000. He prefers the online tools of the ICGC because they offer more scales than Google Maps. G10_P1 wants to have the maximum detail. G10_P2 uses all scales available in Google Maps and zooms in to Street View when needed. G10_P4 likes to see detail at street or even house level.

Name of module[44]	Description as a conclusion of the Delphi study	Input made up of interview group G9 and G10 and participants by email G7 and G14
Type of product	Undergraduate teachers and students and the general public would both prefer to use raster maps, with orthophotos as the second choice. Commercial users would prefer vector maps as second choice. Abbreviated module value: Raster maps, orthophoto and vector maps	G7 uses base maps, topographic maps and orthophotos. G14 is using orthophotos and thinks it is necessary to keep them even though the raw photo would be available too. G9_P1 uses maps to situate land houses of families for genealogical studies, because the map shows detailed place names of individual country houses. G9_P1 and G9_P3 use the official catalogue of place names that also informs of former names of streets and places. G9_P2 and G9_P3 mention occasional use of cadastral data. G9_P3 wants to find owner names related to addresses. G9_P1 and G9_P3 use family registries of the church. G9_P1 and G9_P3 do not know what orthorectification is and none of the participants of G9 are worried about distortion present in raw aerial photographs. G9_P2 knows the implication of projections but says he works with small townships and geographically reduced spaces, where distortions affect little. They would conform to any type of aerial photograph. G10_P3 is interested in maps and photographs, anything that depicts the past. G10_P4 searched for images, aerial images and maps to illustrate new publications. G10_P1 and G10_P4 think it would be desirable to have images of faces of buildings, as they were long ago, accessible similarly to Street View. G10_P2 wishes that all place names of old maps were searchable.
Acquisition interval	The maximum acquisition interval for the general public is estimated to be 5 years and for teachers and students 10 years. Commercial users need all updates of legacy geodata. Abbreviated module value: All updates needed to suit the subgroup commercial users	G9_P1 and G9_P3 agree that for their research a snapshot every 5 years would be enough. G9_P3 says that the frequency depends on the evolution of the townships. For G10_P2, a snapshot every 5 years would be sufficient. G10_P3 would prefer a yearly interval or even more regular to see at which time of year some process started. G10_P1 on one hand says that the more information is available the better; on the other hand, states that having a snapshot every 10 years would be great. He also says that the further back in time, the bigger can be the interval between snapshots.

Name of module[44]	Description as a conclusion of the Delphi study	Input made up of interview group G9 and G10 and participants by email G7 and G14
Significance of the user group	General public, with all related user groups, makes up for 41.5% of all access to legacy geodata. They are infrequent users who demand small or large amounts of data. Commercial users as well as General public are in the middle and Undergraduate teachers and students in the lowest range of user groups concerning their negative impact on economy, if legacy geodata were missing.	G7 consults legacy data up to twice a month. His search is more intense and covers more sources when he does not easily find the answer to his question. G14 considers himself a frequent user of legacy data. For each project, he has to consult legacy data several times. G9_P2 consults and georeferences a map every week. He works on a georeference, on average, for one and a half hours. He states that he is currently unemployed and would not have so much time if he was working. G9_P3 does not consider herself a frequent user, but as a professional genealogist is dependent on the availability of historic information. G9_P1 thinks he consults maybe once a month. All participants of G9 speak of casual map findings that have captured their interest for a moment. G9_P2 consulted German tourist maps from the 1930s he found on paper. G10_P3 and G10_P4 think that they lose themselves in historic information about 3 to 5 times a year. G10_P4 consulted old maps for reproduction in for-profit endeavours. G10_P2 searches background information on a potential client every month, but he does not mention explicitly that he searches for historic geographic information.
File formats	No user specific predictions about file formats were made by the Delphi experts.	G9_P2 downloads PDFs, because the maps he needs are only available in this format, and transfers them to JPG. G9_P1 downloads JPGs. G9_P3 does not know what a vector format is. G9_P2 relativizes the utility of vector topographic data for historic place name searches because the orthography or the whole names change. G10_P1 occasionally downloads a map in JPG format. G10_P3 likes PDF format. G10_P2 thinks that SVG (scalable vector graphics) would be useful, but also mentions that he is OK with any format as long as it appears in the browser.

Table 23: Characteristics of the user groups clustered around the general public.

Additional relevant comments: G9_P2 consults maps online that he would not consult if he had to go to an archive. G9_P2 and G9_P1 agree that the georeference is very important in photographs. For G9_P3 it is important to study historic geographic information about places she can visit when needed. G9_P1 sometimes uses Wikipedia to find maps.

5.2.1.2 Profiles of Catalonian legacy geodata users

Here we will explain where the Delphi predictions are concordant with or divergent from the picture that the current users presented to us.

The user profiles that emerged from the interviews reflect the current state of legacy geodata use and needs in Catalonia. It is known that user interest and research trends arise and disappear. In addition, technological aspects evolve quickly and change the context of use. Therefore, a certain perspective into the future development of user knowledge and the technological environment has been integrated into the Delphi study and assists in interpreting differences with the predicted profiles. Reasons for divergent modules are given at the end of each profile. The following set of profiles offers an example of how the projected future user profiles could be applied to a specific case – the ICGC and the Catalonian context.

5.2.1.2.1 Geographers

The following is the resulting profile of current users in the cluster around geographers.

Name of module[45]	Synthesis of user input	Abbreviated module value
Interaction type	Even though we did not come across many use cases of data exploring in the sense of MacEachren & Kraak (1997), most projects led by geographers involve amending data or creating new vector products.	Query
Time range	Most encountered projects reach over a very large time span. Some projects are limited by the lack of older data, other projects resort to alternative information sources (less precise or harder to obtain) to cover the time frames for which maps and aerial photographs are not available.	Up to all data
GIS knowledge	Several of the interviewed geographers did not use GIS, but all collaborated with people who do so. Usually one person who has a command of the technology in a working group is enough.	High

[45] For definitions of the modules, please see chapter 5.1.1

Scales	Users expressed the need for zooming capability. Current legacy data older than 60 years is available at specific scales only. Geographers extract information from various sources and scales to harmonize it on the scale they need. Currently technological limits do not permit generalizing high resolution images or heavy data loads on the fly. This would be necessary when only large scale information is available and an overview of the area is needed. Nevertheless, it is expected that computing power will continue to increase. Some geographers mentioned that they know how to generalize smaller scale maps out of larger scales.	Needs several scales
Type of product	We cannot confirm high use of databases and vector data with the geographer sample we took. Most of the interviewed geographers mentioned use of aerial photographs, while only a few saw utility in topographic vector maps. Desired features of a topographic map are the place names, a forest and cropland distinction and the transportation layer.	Currently: Aerial photographs; use of databases is likely to increase
Acquisition interval	Even if many geographers would conform to a snapshot interval of 5 years, none of them is really happy with it. Several reasons for desiring increased intervals were given: a) If information for a new layer is collected for the first time, the snapshot frequency should make it clear when information was collected and in which geographic area. b) If a topic is evolving at a high speed, such as the vegetation during the year. c) If important related data is available at shorter intervals, such as annual statistical data. The idea that the time interval between two snapshots could increase with the data age was accepted. One geographer mentioned a cut in the interval frequency 100 years after data creation.	5 years or more
Significance of the user group	Most topics that geographers study have a time dimension, which is why legacy geodata is highly requested. Processes over time are of such importance that students are instructed in the use of legacy data. Geographers' involvement in governmental problem solving would make the disappearance of historic geodata immediately noticeable to society.	High
File formats	Geographers use a variety of vector and raster file formats. For sharing vector data Shapefile is widely used. They are aware of the need for and do file format transformations, but ask for data to stay in its 'original type': vector or raster.	No dominant format

Conclusions: Geographers

The picture that was presented to us by the current geographers corresponds in most modules to the Delphi prediction with one exception: the type of product preferred. There are several reasons for the predominant use of aerial photographs among the interviewed geographers. First, for physical geography the detail of an aerial photograph is important. Because most participants in the geographers cluster came from physical geography, the use of aerial photographs looms predominant. Second, most old maps are not available as vector data. Therefore, when geographers currently use old maps they do not use geodatabases. But the desire to have thematic data in layer format was expressed. Third, some of the land-use types in which geographers of human geography are interested are not reflected in topographic maps; neither are many geophysics topics. If historic maps were in vector database formats, geographers would probably query data by topic independent of the source (e.g., query cadastral data instead of topography or query their own production). Databases will become more useful with the integration of various information sources, such as the geographers' own production or the historic dimension of the gazetteer. Data integration is in process, and newer topographic legacy data have vector formats which is why we believe that the Delphi group made a correct prediction.

Because the time interval between data sets (maps or aerial photographs) is larger than five years for most currently available data from the mid-20th century and older, current work practices would not be severely impacted by a five-year interval. Nevertheless, several geographers have concrete ideas of what processes they would study if annual or monthly data were available. This shows the potential of a high time resolution and therefore the five-year interval is not a recommendation but a regrettable necessity when resources are lacking.

5.2.1.2.2 Historians

The following is the resulting profile of current users in the cluster around historians.

Name of module[46]	Synthesis of user input	Abbreviated module value
Interaction type	Most historians want to visualize information on a map and consult related documents. Of the participants, only one wanted to query geodata. The comparisons one historian mentioned are done currently on paper, but could in the future be executed as a query if data were digital and the interested historian had enough knowledge in geographic information exploration.	Consult and query
Time range	Participants confirmed that their interest in legacy geodata range from most recent to as far back in time as possible.	Up to all data
GIS knowledge	Of the five participants, one confirmed that he had GIS knowledge and another had knowledge in CAD.	Medium to low
Scales	All interviewed historians needed large scales or excerpts from all scales.	Need large scales
Type of product	There is no clear dominance of a type of product. Especially when cartography as a discipline is the object of study, any information source type is of interest to historians. Any individual source can also be the object of study.	All data sources as equal
Acquisition interval	The type of study undertaken by the interviewed historians does not really justify a high frequency of acquisition. Nevertheless, participants wish to have a higher time resolution than 10 years. As a positive effect, the processes they document could be dated more precisely.	10 years
Significance of the user group	The amount of complementary information needed, in addition to maps or aerial photographs, is very high for historians. Their use of actual geographic information is secondary, with the exception of those who study cartography as a discipline.	Medium
File format	Interviewed historians mentioned file formats only related to raster data.	TIFF and JPG for raster data

Conclusions: historians

There are some differences between the characteristics of historians as predicted by the Delphi experts and the ones we encountered in the interviews for two of the modules: scales and the type of product.

Scales: None of the five historians expressed the need for small scales, although this could be due to the small number of interviewed historians. One historian defended small scales

[46] For definitions of the modules, please see the beginning of chapter 5.1.1

because he could imagine that colleagues would need them. Two historians study the context of maps and would need all scales. But they could do their type of research with excerpts of the data if it is accompanied by sufficient metadata or related information such as algorithms that help to generate the data, technologies that are used to render and distribute it, etc.

Type of product: The Delphi experts had predicted users in this cluster would prefer raster maps and vector databases. If we consider that current legacy maps are often distributed as digital raster copies, the Delphi prediction looks more like a description of the current user than a projection into the future. The way the question about the preferred type of product was asked to the Delphi experts allowed them to choose only one type of product per user group. Variety emerged when the experts did not agree. In contrast, the focus group participants mentioned several products they use, but did not necessarily inform about their preferred product.

5.2.1.2.3 Urban planners

The following is the resulting profile of current users in the cluster around urban planners.

Name of module[47]	Synthesis of user input	Abbreviated module value
Interaction type	Even though consultation of aerial photographs is widespread among urban planners, all of the interviewed groups also query and amend geodata.	Consult and query
Time range	Tasks related to urban planning and legal support for ownership need data from the last 50 to 100 years. In addition, urban planners as representatives of the local authority collaborate with cultural endeavours that investigate back to Roman times.	50-100 years
GIS knowledge	All participants in this category know how to use GIS.	High
Scales	Urban planners are responsible for a limited area that is generally smaller than what would be studied by geographers. Their GIS knowledge allows them to recreate small scale maps from larger scales. Nevertheless, small scales have to be preserved when large scales do not cover the whole territory.	Needs at least one large scale of the complete territory.
Type of product	Raster maps and photographs are the most available; however urban planners would like to have vector data as	Raster maps,

[47] For definitions of the modules, please see the beginning of chapter 5.1.1

	well. When it comes to aerial photographs, all mention the orthorectified version. Where geodata is missing, other sources of information come into play, such as construction permits or notarial documents, where the same information has to be extracted from text.	vector data and orthophotos
Acquisition interval	The frequency of acquisition is related to the timespan that urban planners use. For the most recent years and as far back as the oldest currently valid property register, urban planners need to track all changes in cadastral data. For other disputes and lawsuits, it might be necessary to keep map data as old as the public infrastructure in town. The urban planners in this study did not express preferences about time intervals after this period.	Yearly
Significance of the user group	Several urban planners we interviewed mentioned that they systematically acquire geodata from the ICGC, including older flights. Nevertheless, consultation usually concerns a very small geographic location and is rather spaced in time.	Medium
File format	Almost any vector format serves the interviewed urban planners. Shapefile is widely accepted. MrSID outweights TIFF in working environments of urban planners.	Shapefile and MrSID

Conclusions: Urban planners

There are significant differences between the characteristics of the urban planners predicted by the Delphi experts and the ones we encountered in the interviews. Differences include the following modules: interaction type, GIS knowledge, scales, type of product and acquisition interval.

Interaction type: Part of this divergent image might be explained by the size of municipalities that agreed to be interviewed. The smallest township from which urban planners were interviewed had 97,000 inhabitants. This is the 11th most populated of 948 municipalities in Catalonia. Services that exist in populous towns might be more developed than in the many smaller municipalities. It is possible that many municipalities unite functions given to urban planners and other public servants in one person, while at the participating municipalities most often a special department and several dedicated people were in charge of geo-related information.

GIS knowledge: The Delphi experts may have made an assumption for all members of policy makers and lawyers. We interviewed only those who actually handle geodata and did not find lawyers among them. The high number of legal affairs handled by the urban planners led us to conclude that lawyers commission authorities when they need proof that includes geodata. Therefore, the interviewed urban planners are not representative of the total of policy makers and lawyers.

Scales: Depending on the authority and tasks of the policy makers, they might need smaller or larger scales. The statement that urban planners could recreate small scale maps from large scales might also be adulterated by the size of the municipalities interviewed. Indeed, even if urban planners in small municipalities have the GIS knowledge to do so, they might not have the technology or infrastructure. A system of collaboration and assistance between the ICGC and municipalities compensates for lacking resources.

Type of product: In relation to the type of product, we can reinforce the argument already mentioned for historians. If we consider that most legacy maps are distributed as digital raster copies, the Delphi prediction looks more like a description of the type of product used by current urban planners. With the advent of legacy geodata available in vector format and the increased usability of this type of data through internet navigators and freely distributed GIS, it is most probable that future urban planners will use geodatabases more frequently and intensively.

Acquisition interval: To resolve legal issues precisely, urban planners must obtain legacy geodata of each state in the past. This need usually ceases for data older than 50 to 100 years. Differences in this module value might come from Delphi experts thinking of data use older than this. Indeed, we detected decreasing need of snapshots of older date among the interviewed urban planners.

5.2.1.2.4 General public

The following is the resulting profile of current users in the cluster around the general public.

Name of module[48]	Synthesis of user input	Abbreviated module value
Interaction type	Only one user queried data. Some of the use cases could have been simplified with vector data if the user additionally had GIS knowledge.	Consult
Time range	The curiosity of the general public covers all time spans.	All years
GIS	Only one of the questioned participants knew how to serve a	Low

[48] For definitions of the modules, please see the beginning of chapter 5.1.1

knowledge	GIS.	
Scales	When people approach legacy geodata with a precise project and idea, they also need a precise scale. Most of them could not generalize a small-scale map from a large scale. Those users that just wander through available data would not be affected by the elimination of small scales.	Can work with large scale maps
Type of product	The general public uses all kinds of information sources: raster maps, aerial photographs, cadastral data and gazetteers. The use of raster maps is popular for their toponymy.	All kinds of products
Acquisition interval	The participants would content themselves with an interval of 5 to 10 years, but clearly desire a higher interval. The thematic maps they search can only satisfy their curiosity when a higher frequency is chosen.	5 to 10 years for older data
Significance of the user group	Out of 8 participants, 3 consult rather infrequently and the others rather frequently (once a month or more). Only one participant’s income depends on the availability of historic maps. The other commercially oriented participant could find desired information in other sources.	Medium
File formats	All members of the general public handle JPGs and PDFs. They are OK with any format as long as it appears in the browser	JPG and PDF

Conclusions: General public

There are slight differences between the characteristics of the general public predicted by the Delphi experts and the ones we encountered in the interviews. Differences include the interaction type, type of product, scale and acquisition interval.

Interaction type: The Delphi experts predicted that members of the general public query and consult data; however, the dominant interaction type among the interviewed members is ‘consult’. Considering the advent of technologies that ease online distribution and access to vector data and the availability of user-friendly mapping tools, we can predict that in the future the general public will also query data. Indeed, the Delphi experts predicted that at the latest, in 2034, 80% of the implemented GIS will be web-enabled (Delphi group one, round two, question 22). Many Delphi experts think that in the same period up to 50% of the general public might become familiar with querying vector geographic data (Delphi group one, round two, question 19).

Type of product: Two types of products mentioned by the interviewed users were not explicit options in the Delphi survey: the cadastre and the gazetteer. Therefore, the Delphi participants did not express themselves on these products. A cadastre can be consulted as vector data or as a raster map for older editions, which is why it falls under either of those categories that were offered to the Delphi experts. Where the cadastre is consulted for its place name information, it would be helpful for the user to have this information in a structure that can be queried additionally to the raster image (as gazetteer) or replacing the raster image (as vector data with toponymy). The Catalan gazetteer has several forms. In one edition, place names of each region are accompanied by a map and available as raster images. Another edition is a textual database where place names are related to coordinates.

Scales: we discovered two types of users among general public; both need different scales. There are those who approach geospatial information with a specific goal or interest and those who wander through available information with the sole intention to discover something interesting. For the first group, a specific scale or resolution would suit best. The second group has no need of a specific scale.

Acquisition interval: The acquisition interval current users would like to work with corresponds to what Delphi experts have predicted, with the exception of commercial users who are said to need all updates. Considering that commercial users are interested in legacy data up to 40 years old that might be available from a data producer, a long-term archive can stick with the 5 to 10-year interval predicted by the Delphi and confirmed by the interviewed users to minimally satisfy requests from the general public.

5.2.1.3 Limitations of the user profiles

Many interviewed users identified themselves with several user types because their backgrounds are different from their current occupations. Some leisure users had studied geography or history, some geographers are dedicated to history and some urban planners are participating in academic research. This made it difficult to obtain clear profile pictures. Given answers could not always be clearly assigned to one task or another. The negative effect of several users identifying with various profiles was partially offset by clustering the users into groups. Where a participant identified with user groups belonging to the same cluster, it did not matter that he or she had a background in areas differing from the one currently reflected by his or her job profile. This was, for example, the case of a leisure user who turned her use into a business, because leisure use and commercial use are unified in the profile of the general public. The problem existed when the profiles were part of different clusters. Therefore, for some modules the user profiles might not show as much variation between them as they would have with ‘pure’ profiles.

Geographers and urban planners are the ones that most identify with each other’s profiles. Many urban planners have studied geography, geoengineering or a similar subject. Because of this shared background some of their module values, such as GIS knowledge, are similar. Others, such as the time range of data they are interested in and the type of products they use, are very different because they have different tasks.

The Delphi experts predicted the characteristics of western culture users (broadly, Europe and North America), and current users were interviewed in Catalonia, so there is a certain regional difference to expect between the two results. Regional differences might be reflected in professional culture (preference for certain topics and therefore the sources, data age range and scales used to answer related questions), and penetration of technological advances that influence file formats and GIS knowledge. Some differences might come from the fact that the Delphi opinion is a synthesis of the description of various user types, while the interviewed users came from a relatively homogeneous background within the same user clusters. We had intended that the Delphi experts in group one would predict the development of user characteristics in the next 10 years. This should result in a picture of future users that could be compared to the current picture of legacy geodata users from the focus group interviews. This failed in the sense that very few experts expressed themselves on possible changes of user characteristics. Experts described user behaviour and needs but were reluctant to give their opinions on the development of those in the next 10 years. Those experts, who did so, stated that they did not expect much change, so that we had to assume that the predicted profiles would resemble the current users. The proposed 10 years might not have been enough for experts to feel there would be differences in studied topics and technological use. However, a Delphi expert mentions that user interest will turn towards recently made available data. The invitation (to group two) to predict change in technologies was followed. Thanks to the predictions of this group and observation of the ICGCs processes, we can distinguish the following trends:

- Every year new vector data will be superseded and will be made available by the ICGC.
- People consulting legacy geographic information today already wish to query and will learn how to do so.
- Vector-related software will be included in the browser or the service will be on the provider side, so consumers will not necessarily be aware of it.

The availability of vector data and the ability of users will intensify the use of vector formats and influence the modules type of interaction, product type and GIS knowledge. Finally, there are differences between the Delphi predictions and the picture drawn by the current users conditioned by the way the question was asked. The module values of the acquisition interval predicted by the Delphi experts show the tendency to larger intervals because Delphi experts were asked for the maximum time interval with which a user could still work while users tended to bring up the ideal interval for their purpose.

5.2.2 Discussion of user-informed preservation of geodata

We gained insights through the user interviews that can inform appraisal and archival services. While users expressed themselves on their use of legacy geodata the topics of the modules and other issues came up or were intentionally addressed by the interviewer. The frequency with which these were mentioned, the expressed emotions and the emphasis the user put into his or her comments were used to evaluate the appraisal criteria given in chapter 3.2.1 and to discover eventual additional criteria. We discuss these further in the following sections, and propose for each issue the ideal appraisal and selection measure from the point of view of the four user communities.

5.2.2.1 User-informed appraisal

Reactions on selection and potential elimination of data (as part of appraisal) were expected to range from critical to openly negative, because preserving everything forever is a known user desire (Albani, Guarino, & Leone, 2011). Indeed, most interview partners expressed negative feelings ranging from surprise to open rejection of the idea, with the exception of some participants representing the general public; they seemed to accept elimination of certain data more readily. When the possibility of appraisal was accepted, some participants from the general public and geographers came up with service level ideas to adjust cost or effort levels.

The first criterion for preservation discussed in chapter 10.1 is the value of data. Users did not mention value explicitly. Users expressed value through negative emotions when faced with a threat of losing data. Looked at in this light, all data from the ICGC are valuable because users seem to wish to preserve it all, even though they have never needed it and would not even know how to render or interpret it. Therefore, these expressions of value are of little help for appraisal. We detected what users value by listening to what they invest resources in obtaining and what their reasons are for finally using a data set.

For municipalities, legacy geodata has informational value first of all, because they use it for their work (e.g., determining house numbers or changing street names back to the 'original'). In some cases, it has legal value (e.g., for the urban planner who had to date the construction of a building and when a company sued the local authority for having cut an important feeder; the state of the official geodata had to be restored to prove inexistence of this information). In the second case, the economic value of the data is obvious. In the previous examples, availability of legacy geodata can speed up the work of the employees, such that the township gains in human time that can be invested on other projects, which is indirectly also an economic value. Geographers use legacy geodata to build and test models and to describe landscapes and social, economic and physical processes in a region. Maps and aerial photographs are equally valuable to them and drawn on in research projects, which is why we can say that the production of the ICGC also has research value. Many historians reconstruct situations that can be visualized on a map. Examples are the core township centres in the vinery region and the creation of a map of the town in Roman and medieval times. This type of task will be easier in the future if we keep geodata, especially vector data, available and useable for centuries. We conclude that geodata also has historic value. The interest that the general public has in discovering change through photographs and maps and sharing it with others shows the social and cultural value of the data. Urban planners, geographers and historians also participate in cultural endeavours where legacy data is used. Many general public users wander through historic information about their town or any other region of their interest, while the genealogist in the study makes a living out of the social curiosity of her clients. In one data set we can even identify intrinsic value in the eyes of the users, because it is the first of its kind. The use and reasons for use the interview partners explained to us let us conclude that geodata from map production have all kinds of value to users.

Feasibility:

Feasibility is an appraisal criterion that goes against the perception of most users that preservation is resolved easily with cheaper storage media or that the only preservation problems are of a technological nature. Users that have experienced preservation challenges in their institutions, such as access problems to obsolete file formats or shortage in storage space, recognize that questions about feasibility must be asked. Feasibility is an appraisal criterion that must be applied for the users' own benefit, given that, if the archive is not sustainable, all data might be lost and there would not be anything to access.

Usability:

Application of the usability criterion favours the user. G4_P1 has her own definition of useable data. She states that data recovered from the archive many years from now should at least be readable by the computer, then combinable with recent data. It should automatically overlap with maps of the time, and if this cannot be the case, due to differing projection or georeference systems, there should be enough metadata to make the conversion. She brings up aspects of understandability (how the projection works) and renderability. We will look at these two aspects, as well as authenticity, reliability and integrity already mentioned in chapter 3.2.1.2.2.

Understandability: Understandability involves having intelligent data and sufficient metadata. G10_P2 finds that maps are only useful when they are accessible as intelligent data (vectorised). Various users (G3_P2, G2_P2, S3) are worried about keeping the attributes of features; it is important to know what a symbol meant. It is also important to keep the relationships between an earlier way of representing data to a newer way, in the same time series (e.g., what before was red is now represented in blue). Users expect data to be easy to combine with existing current data, e.g., Google Maps or Google Earth, to understand which area the data is concerned with. This is given with georeferenced files for which the coordinate system and the projection is known. On the one hand, even though users might not know what a georeference is, they require the feature and have a hard time situating data when a georeference is missing. On the other hand, most users do not worry about the projection. Some work on a very large scale where projection does not matter, others just consult data or use them in a context where precision does not matter, e.g., for a historic publication.

Renderability: Renderable data is compatible with current hardware and software. Members of the general public may expect files to open automatically in their navigators. G10_P1, for example, is not completely aware which program opens the maps he downloads. Other users choose to download data in a file format they know their software can handle, if there is a choice. Problems with renderability because the processing power of the users' environment is not high enough for the data can be estimated for the current environment but not predicted for the future. Some users in this study are aware that there might be a rendering problem with legacy data in the future, due to obsolescence of hardware or software, but they do not connect it to the effort and cost needed to keep data renderable.

Authenticity, reliability and integrity: Only two users in the study are worried about data reliability and authenticity. One user theorises about the feasibility of quality checks on regenerated data. Another is concerned to know the initial purpose of the map, because maps can reflect propaganda or wishful thinking. This speaks to the importance of knowing who the initiator or contracting body of the map is or was and to what purpose it was created. None of the remaining interviewed users mentioned authenticity, reliability or integrity. From this, we conclude that they trust the sources they consult or there is enough information for them to judge the reliability of the data so that they do not have to worry about it.

Organisational focus:

The organisational focus concerns the collection development policy or the mission statement. Users in the internet age do not care who holds the data, they want to find everything in the same place. The trip to the physical archive is seen as a burden. A user mentions that 'his' archive can stop purchasing maps from overseas when they are available online. Therefore, accessibility decisions of other archives can prompt changes in one's own collection policy. Collection development policies might exclude parts of data sets when they cover a territory larger than the one projected in the policy. Such data sets are at risk of losing completeness. Archives should collaborate to strive towards preserving such large data sets completely and maintaining their accessibility under similar conditions between archives. Ideally for the user would be having a single access point (Pérez et al., 2013). In a world of common catalogues and easily accessible content, combinability of data with existing collections should be checked against data sets beyond institutional holdings.

Potential future use:

Several factors influence future use in the eye of the user: Homogeneity, the spatial and temporal resolution and the spatial and temporal coverage. In general, we can say the higher the temporal and spatial resolution and the larger the temporal and spatial coverage the more use cases are imaginable with the data. Data sets that are detailed and large in both senses are very big. A limitation for their use is the user's IT environment, which might not handle such objects, but this should not influence the appraisal decision of the archive. A service that helps to split and obtain parts of the data set should be considered, however.

Whether or not a data set can be reused to answer a specific question depends on the size and extent of the studied phenomenon. Spatial resolution or scale and coverage have the same impact on use of current data as on legacy data and will not be further analysed.

Eleven out of thirteen users who participated in the questionnaire point out that homogeneity is 'somewhat' to 'very important' for their comparison of maps. Homogeneity is higher in data within the same time series because semantics are likely to stay the same. Most interviewed users express the desire to compare data over a larger time frame than currently available and at a higher time resolution. The following two figures show the relation between the time span, time resolution and the phenomenon users want to study.

Time series that span a long time frame are more susceptible for future use because they can answer more questions. In the example below, data set HTR allows studying various stages of process 1 and 2 but does not cover process 4. If the time span were some years longer, scientists could also analyse process 4. Apart from the time span, the time resolution influences future use. Even though process 3 occurs at a time covered by data sets HTR and LTR it cannot be observed. Data with a higher time resolution would be needed to observe process 3. High time resolution allows comparing more stages of processes (10 stages of process 1 for data set HTR in Figure 9 against 2 stages for data set LTR in Figure 10).

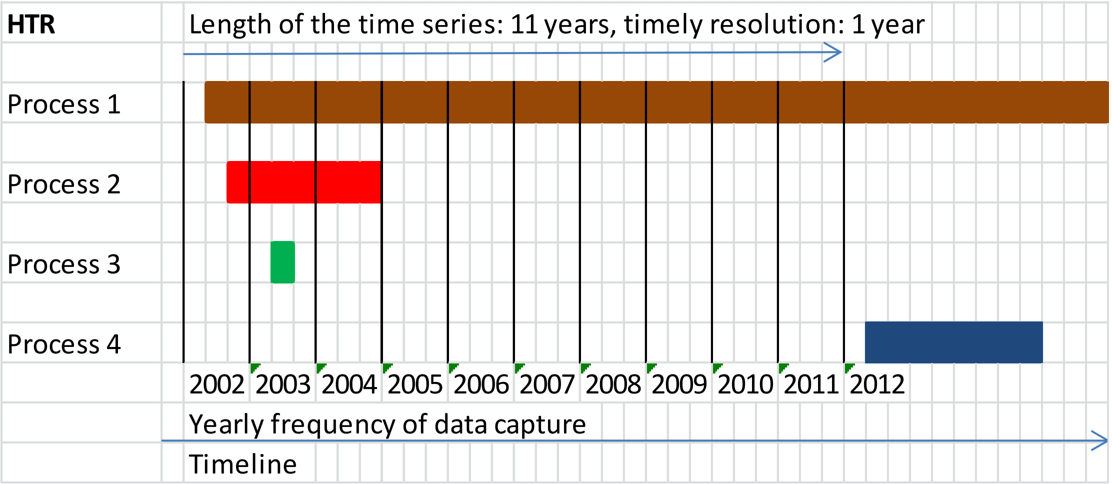

Figure 9: Data set HTR with a high time resolution of one year shows many possibilities for dating and change analysis for four processes of different length

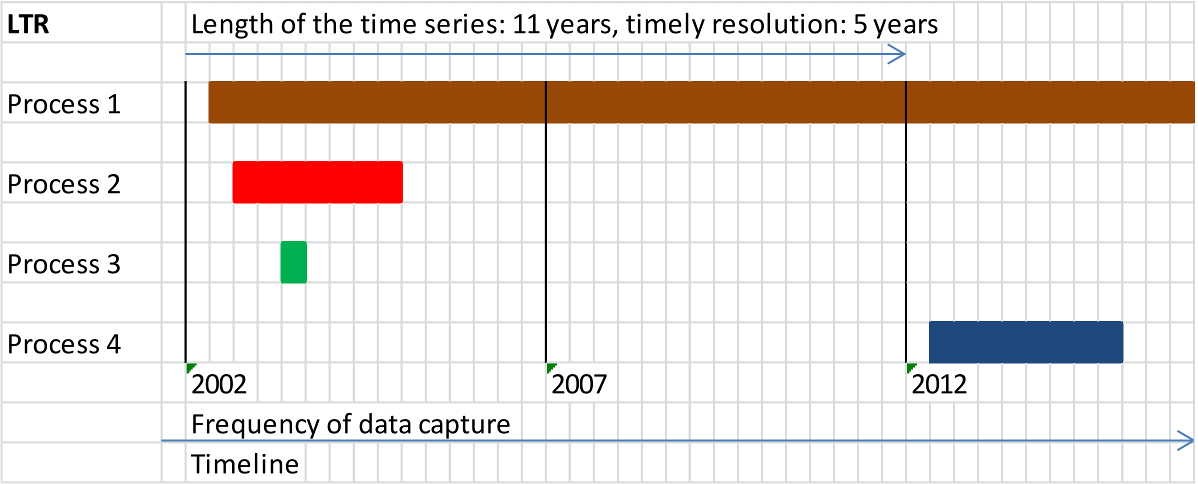

Figure 10: Data set LTR with a low time resolution of five years shows the possibilities for dating and change analysis of only one process.

The time resolution is highly related with the acquisition interval of the archive (see chapter 5.2.2.2). Apart from change analysis we found several users wanting to date events from the past. The time resolution of a data set determines how precise dating can be. Data set HTR shows that the beginning of process 1 and 2 was sometime in 2002. From data set LTR we can only determine the beginning of process 1 as laying between 2002 and 2007. The precision to which a process must be dated depends on the subject. Here are some examples: for legal reasons the construction year of a building has to be found; for a biological research study, regional variation of the flowering period of a plant needs dating in terms of weeks; and to date the change in the course of a river, the century or eras might be enough. Larger time spans increase the chance that the data covers the desired process while a higher interval allows a more precise dating or a more detailed change analysis.

Scientists are aware that their field of study is very specific and their production might not be reused very often. This makes preservation of their production secondary compared to basic geodata, such as data produced by the ICGC. The study users expressed this opinion in two ways. First, some geographers mention that their map production is an interpretation and that new knowledge about the subject (by them or others) will result in different or even better interpretations and representations (G6_P3 and G11_P1). Second, another user says that he would not want all data interpretations that were made in the past to be available, as there must be 'something left' to research.

Interviews also show that use might shift due to data source alternatives. Some of the interviewed users choose Google Earth or Google Maps to view legacy geodata. It is unclear if this will increase use of the sources at the ICGC, because geodata use becomes more widespread, or if it will decrease its use, because the alternative tools might be more accessible or user-friendly.

Data quality:

We will look at the three aspects of data quality mentioned in chapter 3.2.1.2.5: spatial accuracy, resolution and measurement certainty; quality as completeness of descriptive and technical metadata; and material quality. We must take into account that geodata users have their own concept of quality: users check if data is 'fit for use' (Boin & Hunter, 2006), which corresponds to our criterion of usability. Boin and Hunter show that metadata fields such as lineage, accuracy and completeness are rarely consulted or are found confusing and therefore do not influence the decision about data being fit for use.

Our interviews confirmed that users trust in data accuracy rather than check it. They express their need for accuracy when elimination of the orthophoto is mentioned and they fear the risk of having to regenerate it. The accuracy of aerial photographs lies in the quality of the orthorectification. Doubts on accuracy occurred related to use of old maps, while general trust was given to recent data and were related to remote sensed data. Several users, including a geographer and a historian, perceive old documents as equal to bad quality. G1_P2 and G1_P1 say that the accuracy of maps and quality of georeferences are questionable for old maps, which supposes a lot of work to make data sets fit each other. A historian says that the plan data she uses is rarely accurate, which increases the time she invests in the project, but she does not refer to topographic maps. None of the interviewed users mentioned searching for measurement instructions, but finding and improving the projection of old maps is a research issue (Baiocchi & Lelo, 2010; Bayer, 2016). This laborious work can be avoided for future researchers if we document the projection of our current maps.

Worries about resolution quality in data capture (e.g., for aerial photographs and scans of analogue information, such as paper maps) are mentioned explicitly. Nevertheless, whether or not a resolution is good enough depends on the specific use case. For example, S1 and G5_P1 say that reading inscriptions on maps is sometimes difficult at the scanned resolution.

Users worry about metadata. This is exemplified by two users (S3 and S7) who state that it is difficult to interpret old maps if they have lost the legend or others who try to identify the author of a map. Nevertheless, users do not care about the completeness of the metadata set as long as they get the elements they need. The archive has to check completeness for the benefit of all users because for some purposes one metadata element is important and for other purposes another element is important. Metadata elements that could be required can be deduced by the way users would like to search: by location, date and topic. Georeference is by far the most desired metadata element by all user groups. The date is also very important to users, to include the year the real-world objects were observed and the date data were interpreted or assembled. For aerial photographs this is the same date, but vector data can be created many years after initial data capture (the photograph is taken). Deferred creation is the case for many thematic maps that stem from interpretation of other sources, as mentioned by the users. Finally, the topic is a desired field that should be searchable.

User comments do not reveal if trust in recent data is given because of knowledge about the data source's (the ICGC's) quality practice and authority or just because the data is recent and it is supposed that they are made with the currently best methods. In the first case, trust in quality should not decrease with data age unless knowledge about the creator's authority gets lost. In the second case, trust would decrease because newer technology would do it better. In any case, users who worry about accuracy compensate for its lack by adapting data to their needs. We did not study which information they rely on in this case, but the presence of technical metadata such as projection, scale and coordinate system for maps should resolve ‘mismatching’ of some data sets, which is one of the reasons why data sets are perceived as being of bad quality (Boin & Hunter, 2006).

Completeness of technical metadata and material quality is something the archive must worry about for the benefit of the user. Most users just expect data to work on their screens.

Data completeness:

We defined data completeness as the presence of auxiliary documents, e.g., measurement instructions for the data set (completeness on the data set level), and the presence of related data, e.g., base maps, corresponding digital terrain models or thematic data (completeness between data sets). Completeness at the data set level can be seen as metadata completeness and is discussed in the previous paragraph.

Some data sets are partitioned between archives because they are too large or because the collection policy of an archive is tailored to the region of its political authority. From the point of view of the archive the data set might be complete, because it covers the whole geographic area in the responsibility of the archive. In the eye of the user who studies a larger area, it is incomplete. The user perceives incompleteness on the data set level, while as for the archive it is an issue of completeness between data sets.

In a world where interdisciplinary research is practiced, data completeness is important not only on the data set level but also between data sets. For many enquiries on the academic and personal level, data of a variety of topics are necessary to get a complete picture of the situation. For some projects, distantly related sources must be used, such as in the case of a user who analysed truck sales to deduce if that region was deforested at a certain time. Users spend a lot of their time searching for the data needed, especially when it is not digitized and distributed over various sources. If data is not found it is created. G2_P2 estimates that 80% to 90% of the overall time spent on an academic cartographic project is spent assembling data. The process of searching for and obtaining data is perceived as a burden in academic, professional and private fields and sometimes a barrier to use for the general public. To be able to offer interdisciplinary data sets, current best practice is to interconnect and merge data catalogues rather than widen individual collection development policies to include new disciplines. In the case of partitioned data sets we would recommend that archives assure related data is preserved and accessible elsewhere through institutional networking.

Uniqueness:

The three aspects of uniqueness we addressed in chapter 3.2.1.2.7 are repeatability, available copies and redundancy and substitutability.

Repeatability: As said before, some processes of map making are repeatable (if source data and algorithms are kept). The interviewed users trust the production of the ICGC, and we can suppose that they would also trust the regeneration of the product by the same institution. Nevertheless, because not all processes are 100% automatic, results would not be identical to the originally experienced product. From user input we deduce that the fact of the products not being identical is no issue as long as the original process is documented.

Available copies: Several users mention having acquired copies of data sets stored locally for easy access. The users' institution is therefore a competitor on user access to such data. In theory, if other archives hold the same data set, its preservation is guaranteed and access and other services are equal, no additional copy must be preserved. Nevertheless, when several archives hold and offer the same data it benefits the users, because of increased findability. Users do not care which institution is behind the offered data when they are accessible, but users have their preferred data sources and would like to access all at the same place. If an archive decides to retire a data set that is available elsewhere, it should direct its users to the other source. Apart from keeping data findable and easing access for the user, the archive might have other reasons to preserve the data set. For example, it might be credited for the number of users attracted by the data set.

Redundancy and substitutability: There are several examples of redundant data – the generalized scale that is contained in a larger scale and the similar data sets (geographic coverage and resolution) produced by other bodies. In theory, a data set that is automatically generated by other data is redundant and could be eliminated. On the one hand, users confirm this because they require the most original data and study the largest scale available about the topic of their interest. On the other hand, users want to zoom out of their local area for orientation and overview. Derived scales can be eliminated as long as zooming is possible. To provide the experience of zooming for the user, the archive has several options:

- Regeneration of the scale on the fly, based on the larger scale data preserved for the same year. This depends on the technology available and would be an ideal solution even for the general public. Technology should resolve the generalization request in an appropriate speed.
- Skipping the eliminated zooming level. The visualization jumps to the next lower resolution which creates less smooth zooming.
- Offering more recent small-scale reference layers. A member of the general public user group mentions this as her solution of choice in the current situation, where for historic maps no set of scales is available. She even makes the effort to switch between visualization tools because the historic data and the reference data are not integrated.

The Spanish National Geographic Institute produces vector data at a 1:25,000 scale that could be used as a substitute for the vector data of the same scale from the ICGC. Most interviewed Catalonian geographers and urban planners work with local data. Some study a phenomenon of their region; one geographer works with Swiss data, because he modelled a general physical phenomenon. The participants of the cluster ‘general public’ use map scales up to 1:10,000 but would like to have more detail. The larger scales that users work with are not produced by the IGN and have no substitute, but for zooming and giving a general overview, the 1:25,000 scale of the IGN could substitute for the data set of the ICGC.

Additional criteria or indicators:

Three urgent issues for the users, usability, quality and completeness, are addressed by existing appraisal criteria, but the most transversal requirement – accessibility – is not satisfactorily covered by feasibility guidelines and will be approached in a separate chapter about service levels (see chapter 5.2.2.3). User comments let us detect two additional arguments for preservation that the analysed appraisal guidelines did not produce in this detail. The interviews made it obvious why the aerial photographs from 1956 are so important to users: they are the oldest available[49] and they have good resolution. Many

[49] At the time of the interviews the photographs from the American flight series A from 1945 were not available in digital form yet.

users would like to use aerial photographs older than this, but either they do not know of their existence, or they consider them to be of poor quality (referring to older flights) or too hard to access (referring to print photographs at the tax office). Being the first of its kind seems to add value, especially combined with accessibility, because it is stimulation for use. Being the first of its kind could be a prioritization criterion for service levels. If the data are really the first of their kind this is an indicator for intrinsic value. If data are only the first accessible versions, they have research and social value as long as they are not overshadowed by older and equally accessible data. This was recently the case when the American flight Series A was scanned and put online. It could be said that being the first of its kind is an argument for originality in the sense of the DANS guidelines.

For the interviewed users, one of the reasons to use the aerial photographs of 1956 is because they show the territory just before a time when urban change started to be significant. Because appraisal for the archive takes place several years after data creation, a criterion such as ‘documents a state previous to or subsequent to an event’ could be imagined. Depending on the type of event, this could be a criterion for social, historic or research value.

5.2.2.1.1 User-informed selection of file formats

Most users like TIFF files for their resolution. However, several users state that TIFF files or other high resolution raster data are not manageable (the files are too big) for their IT environment or that they use other file types because they do not need high resolution. Even though we met geographers that only manage PDFs for their projects, team members managed the vector files. Those who have GIS knowledge work with any vector formats. Most members of the general public are fine with JPG files. In general, users do not worry about file formats as long as the file is supported by their software. Therefore, we can say that the usability or functionality is important to the user. How manageable a file is for the user’s computer depends not only on the file format and resolution but also on the area the user needs. A large area at a high resolution corresponds to a high data volume. Because the user expects a ‘useable’ file he or she rarely will be satisfied with bit-level files that might serve as preservation masters. Users expect the archive to transform files or to hold a useable file available for them. If the archive can choose a file format that is recommended for being a standard and an open format that can also serve as a distribution format, preservation of data in that file format will gain in feasibility, because less transformation will be needed. The extent of standardization of user environments will determine if the archive has to provide one or more file formats of the same data.

5.2.2.1.2 User-informed selection of scales

Current historic geodata (old maps) are generally only available at one specific scale. Elimination of derived scales after a certain grace period would therefore not hamper the way current legacy geodata users work. But we want to expand the possibilities of researchers while minimizing the size of the collection. Eliminating a scale is only an option when it is derived and the conditions of repeatability, redundancy or substitutability are given.

If information had been added to the derived scale that does not figure in the larger scale, the additional information should be preserved as an individual layer. People wanting to regenerate the smaller scale should be provided with these additional layers. Geographers mention that they would create their own set of scales where necessary.

Among historians we find those who study the map or geodata product itself and those who study the history of technology in mapmaking. Those people need to see how the product was made and presented to the public. Therefore, they need at least a sample of each scale and each data product. If some scales are eliminated they need to be able to recreate those scales. They also need access to algorithms and metadata. Regeneration of missing scales might be a service of the archive, but it is to be expected that in an academic environment they will find competent collaborators.

Urban planners in the larger townships, such as those we interviewed, would be able to recreate larger scales from smaller scales. Besides, if data is kept available 50 years before a selection of scales is made, they might never have to recreate older scales.

Activity of the general public would not be limited by the elimination of scales as long as zooming in and out with the remaining scales is still possible. The general public might recreate scales for further publishing of additional data (thematic maps). We would not expect them to generalize large amounts of data, but instead only the area covering their interest. The general public typically uses what is available.

In the future, we could expect a geographic database to hold digital geographic objects only in the most detailed version and to render different scales automatically.

5.2.2.1.3 User-informed selection of products of the production chain

We would like to analyse if preservation of any of the current by-products of map creation is unnecessary. First, we check for substitutability of maps and aerial photographs. We mentioned in chapter 4.1.5 that aerial photographs are an early by-product of map creation. Because of the flow of information between photographs and maps, for many use cases maps and aerial photographs can partially replace each other as information sources. Nevertheless, current legacy data users often cannot choose between maps and photographs when older data is needed because only one option exists. Interviews showed that what participants use now is to a high degree influenced by what is available and the trust that can be put in the accuracy and quality of the information. For example, G5_P3 works with aerial photographs, but would rather use historic maps if such were available for the same area and year. For other cases, such as place names, political borders that are not on aerial photographs or natural phenomenon that are not on maps, the two products cannot be interchanged.

Delphi experts expressed their opinion on whether transfer of information from raster images to vector data could soon be automatic (Delphi group 2, round 2, question 50). Their answers suggest that there is still quite a long way to go, and that the resulting data might not have sufficient quality. It will be easier to extract vector data from rasterised maps than from aerial photographs, because boundaries between features are clearer and an algorithm can be fed with the significance of symbols. It is therefore an option for scanned maps.

If this technology becomes available at the archive, its staff can reassess the need to keep the easy-to-extract layers. Nevertheless, for researchers of the history of map making and users with a need for increased authenticity, regenerated vector data will not suit their purpose, because regenerating such layers will not show vector data exactly as perceived originally when they were still elaborated by humans. Transfer of the information would be better when the source is a raster map than when it is an aerial photograph. Therefore, this technology will eventually help to avoid preservation of vector data when raster representations of that data are available.

For the above reasons, and because users consider both aerial images and maps as end products, we conclude that at present at least one expression of both must be preserved.

As follows, we explain separately if any of the versions of the aerial photograph or any of the versions of the vector or map data can be omitted.

Aerial photograph: Technology and services exist to convert aerial photographs on the fly into orthophotos. One user states that he does so with a tool of the ICGC. Other users mention that it is possible to save the georeference points and process to regenerate orthophotos on demand. Such data can be considered repeatable (criterion of uniqueness is not given). If the whole generation process is automatic, saving it would allow regenerating the exact same orthophoto. With this technology, the archive would not have to preserve the digital negative and the orthorectified photograph, which would save a lot of disk space. Orthorectification is resource intensive for computers. The service should be executed by the archive due to usability requirements of most users. If the digital negative is accessible and georeferenced, we would expect that orthorectification would be rarely requested for older data because most users do not worry about distortion and could use the raw photograph.

Vector objects and map data: given the recommendation of LoC to consider geographic data objects and their representation as two different products that deserve preservation (see chapter 3.2.2.1), we think that a version of each should be preserved.

Vector data objects should be preserved because they represent the promise of potential combinability and exploration with other intelligent digital data. The argument of being the original and source of the other is only secondary. Preservation of the raster version is defended with the argument of usability. The argument for vector data preservation is stronger because data will never lose research potential. Usability issues that make us preserve the raster map might disappear – vector maps might be directly rendered online and arranged in an aesthetic readable way so that no raster maps need to be produced. In addition, if the archive can recreate the visualizations as they were perceived, historians and the general public would be satisfied. Historians are interested in how people interact with the map – who created it and why, what was the social reception of this information, who knew about it and what did they make out of this information. People nowadays interact with the published version of the digital map, which is why historians would prefer us to keep the published product.

If the archive cannot recreate visualizations, samples visualizations should be preserved for usability reasons for the general public and for historic research that studies map making as a subject.

As a conclusion, and to serve all four user profiles, we recommend preserving

- a versioned vector database of the most detailed geographic objects.
- the official vector data that covers the whole territory.
- a raster base map of each official scale for consultation and rendering.
- a sample area of all derived raster maps representing each collection policy. This means that each time the collection policy or the definition of features change a sample has to be taken where these changes are made visible (compare with Figure 3).

We also recommend eliminating the derived geodata vector product that served to render derived raster maps. Geographers can recreate it if needed for query, historians can consult complementary documentation on how it was created or consult the samples, and the general public and urban planners might not require it when it is older than 50 years.

5.2.2.2 User-informed acquisition interval

The archive’s decision about the acquisition interval may be the measure that most restricts or stimulates future research. The acquisition interval determines the final temporal resolution of the preserved data set. Similar to spatial resolution, high time resolution allows answering many more questions (see the criterion ‘potential future use’ in chapter 5.2.2.1). Remembering the current situation; available legacy geospatial information sources of an older age are widely spaced in time and do not have a regular resolution. Legacy data of the last 30 years are available in almost yearly intervals but are not in vector format. Irregular acquisition intervals are suggested for project-based data. Examples are aerial photographs that are triggered by a special event or thematic data collected by researchers.

For regular production, the acquisition interval should take into account the frequency of update. When the acquisition interval can be maintained over a longer period the data set time series would be homogenous in at least one property. Homogeneity is important when data-intensive research, such as research of long time–series data, is performed where there would be lack of time for individual interventions on data sets. Homogeneous acquisition intervals foster some types of explorative research. Short intervals widen the types of patterns that can be detected. Event-triggered acquisition fosters detection of other types of patterns, such as variation of the event itself. We can expect that with the availability of legacy data in vector formats more research will be of an explorative nature.

To increase the research value of maps for interdisciplinary research in the social sciences, an interval equal to demographic and other statistical survey periods is recommended. Because geographers of human geography often use such sources, a yearly snapshot is highly appreciated by those scientists. Because the frequency of map making is unlikely to increase, scientists who need monthly or even higher frequency of data will always have to draw on alternative data sources such as satellite images.

5.2.2.3 User-informed services levels: increasing accessibility and usability

When reviewing the issues and problems users encounter when using legacy geodata we see that all turn around accessibility and usability. All of the following conditions that increase accessibility and usability could be regarded as service levels. Each issue is accompanied by statements of users.

1) Data are digitized or born digital. One historian (G5_P1) says that he intentionally uses only digitized sources, because he does not have the means to digitize himself. Two historians (G5_P1 and G5_P3) agree that searching in paper archives takes a lot of time that is not always available. Even an electronic request can be perceived as a too high a burden.

2) Data are available online. Because being digital is a condition for being available online, both are often used interchangeably or online availability is implied when speaking of the need to have data digitized. Usually large digitization projects have the goal to put at least one copy online. G9_P2 and G9_P3 insist on having things available online for free. G2_P2 says that information on paper is hard to obtain and he sees it as a cultural obligation to make geodata available online.

3) Data have quality descriptive metadata and therefore are findable in search engines or on other popular sites. This is critical, because data that are not found are considered to be nonexistent. When some interviewees say that some data sets do not exist, while other users mention their existence, we can conclude that there is a findability problem. G10_P1 gets the maps he needs from a basic image search in Google, another user from Wikipedia.

4) Data are georeferenced. All users are very concerned with localising the found information. Missing georeferences are by far the most mentioned problem of usability. G3_P2 and his group are georeferencing maps from the local paper archive for the use of other departments on the township. G9_P2 references raster maps in his free time.

5) Data are rendered on a base map. This means data are findable when browsing to the geographic area they cover. This is only possible when the files are georeferenced. This helps all users, but especially the people who cannot read coordinates, such as most members of the general public and the policy makers G3_P2 mentions. G10_P2 would like to have historic maps overlaid in Google Maps.

6) Data are free (no cost). Academic actors and urban planners mention that they buy data. Members of general public positively mention low prices of geodata.

7) Data have no restrictive licence. Licences are often related to cost. Interviewed users are not concerned with licences. One user mentioned obtaining a licence for publishing a map. Two other users deliberate about whether cadastre data is publicly available. Most users do not come across licence issues because either their institutions have acquired the licence or the data they consult is in public domain. The ICGC does not digitize maps younger than 70 years, which is when they become part of the public domain, if they are not of their own production.

G6_P2 mentions various barriers to accessibility – information is not online, is not georeferenced and is not free of charge. For G3_P2, accessibility means that there are geographic access points to archival content, and it is georeferenced and digitized. G6_P1 is an exception. For him it is harder to access digital information, because he was used to the print version for many years. His example though, refers to statistical tables and not to maps.

Several users make statements on the access points the digital archive must have. This was not a question of our research but emerged naturally. Many users, for example G6_P3, had the urge to explain how they would like to search in an archive – by location, either typing a place name or browsing to the place over a map; by year or era; and by theme or topic. Indeed, for many of the encountered research projects a lot of thematic information is needed. In many cases, thematic information is created from sources other than the base map and even from non-geographic sources. The idea of access by location is to click a place on a map and the system would show all documents (geographic or not) available related to the selected place. If maps are mixed with other documents, then the search results, catalogue or inventory should show which archival material contains cartographic elements.

The importance of access to a data set is showed in the American flight series B data set. In the eyes of the users, this data set, had intrinsic value for being the first of its kind, even though series A existed and is 10 years older. The difference was that series A was not digitized and therefore was nonexistent or inaccessible in the opinion of many users. It remains to be seen if the digitization of Series A will steal the limelight of Series B.

5.2.3 Proposed appraisal decisions for the ICGC

We would like to propose a designated community for the ICGC and in conclusion address some of the consequences for selection and appraisal of their production. We could say that the four analysed user groups are already implicitly present in the ICGC mission: Because of its geographic nature, the ICGC's production is the primary data for many geographers. Because of its duty to preserve the cartographic heritage, the ICGC's library collection is an ideal source for historians. As a government body networking and collaborating with other government bodies on all hierarchical levels, it is already a service to urban planners. Finally, its access strategy and free data distribution benefits the public in general. The OAIS reference model, nevertheless, requires that the designated community be defined explicitly. Additionally, the OAIS reference model argues that

[…] information originally intended to be understandable to a particular scientific community may need to be made understandable to the general public. This is likely to mean adding

explanations in support of the Representation Information and the Preservation Description Information, and it can become increasingly difficult to obtain this information over time.

Considering that an institution might change designated communities over time and the disposition of the ICGC to serve data to the general public, we recommend defining the general public as its primary designated community. The general public is very demanding regarding its module value 'GIS knowledge'. Because in general members of the general public have low GIS knowledge, choosing this broad community implies assuring complete data descriptions without jargon to make data understandable to anybody and technical services that minimize manipulation required of users. The general public has module values that induce less demanding needs than other user groups. Therefore, sole consideration of the general public's module values in appraisal decisions would lead to restrictions for other user communities. Adapting appraisal decisions to the general public should not impede data use by other communities, such as would happen with reduction of available information to a raster type of product.

As long as the ICGC archives its own data, it has no difficulty in obtaining technical and descriptive metadata needed for preservation. Because of its heritage function for cartography in Catalonia, it will acquire data from third parties. The ICGC should require sufficient descriptive information on that data and eventually reject it, if no such information is available.

Because creativity is involved in raster map and vector database production (organising and classifying objects and arranging and displaying features) they are considered creative works. Raster maps that are used for online display are considered published. When an online map is displayed directly from the vector data, the underlying data is considered published. Published work is subject to legal deposit under Catalan law.

As follows, we will apply the reflections of the user-informed appraisal recommendations to a part of the ICGC's production, first for maps and then for aerial photographs.

5.2.3.1 Map scale 1:1,000

This is the most detailed scale that is considered a base map, but covers only urban areas. It contains information that is not available in other products of the ICGC. This scale is rendered online directly from vector data and should be preserved as such. For administrative reasons, each update should be kept for 50 years in a way that makes it obvious when which information was added (requirement of urban planners). Because this scale is currently updated in a 4-year interval this interval can be maintained for snapshots. This means that, if during the first 50 years a versioned database was available, data density can be thinned to one snapshot every four years. Versioning should occur at the end of each editing process.

This scale should be checked for redundancy with larger scales at municipalities that create their own map data. If larger scales are available at similar intervals this scale might be reduced after 50 years to the municipalities that do not create their own data. In its duty to preserve the map heritage of Catalonia, the ICGC should acquire larger scale data or make sure they are preserved and made accessible under similar conditions.

5.2.3.2 Map scale 1:5,000

This scale is available as a base map and as a map. Both contain the same information and are considered redundant. Because the base map is official, it must be preserved. The algorithm that is applied to the base map to obtain the map should be preserved as a document that records the ICGC's practice at the time.

All updates of the vector data should be preserved for 50 years, if possible (requirement of urban planners). Otherwise, at least yearly snapshots should be made. Currently the raster map is part of the production chain of this scale because the related vector data cannot yet be rendered online automatically. As long as this is the case, all editions of the raster representation must be preserved and made available online. When future technology is able to render legacy vector data online, sample raster maps of this scale will serve as illustrations of the technology in use at the time of their production (requirement of some historians). When raster maps cease to be part of the production chain, they do not have to be generated for the archive unless preservation of the vector data is not feasible.

After 50 years, yearly vector data versions should be preserved to allow interdisciplinary research with the social sciences and other sciences with a yearly interval of data capture. Time resolution of raster representations can be lowered to a five-year interval, or if technology allows automatic rendering of the vector data, reduced to the samples mentioned in the previous paragraph.

5.2.3.3 Map scale 1:10,000

This scale is derived from the 1:5,000 scale. It exists as a map only. We recommend keeping the data available for 10 years and thereafter preserving a data set that is representative of each process or policy that has an effect on how data is displayed. As long as the data are distributed as raster graphics over the internet, they should also be preserved as raster graphics, for authenticity reasons.

Vector data will not need to be preserved, because they derive from the larger scale data. But the algorithm for generalising should be preserved, for the sake of the history of map making (requirement of historians). If there is thematic data of this scale, the corresponding reference data can be regenerated from the preserved vector data at the 1:5,000 scale.

5.2.3.4 Map scale 1:25,000

This scale is available as a base map and as a map. The map contains additional layers and classes not present in the base map. Because the base map is official, it has to be preserved. The additional layers and layers with additional classes should be preserved as well. Because the requirement of urban planners to keep every update is satisfied with the larger scale map, it is not necessary to apply the requirement to the smaller scale maps. We recommend ingesting this vector data in the archive each time a new edition of a tile is published and preserving it indefinitely.

Currently the raster map is part of the production chain of this scale because the related vector data cannot yet be rendered online automatically. As long as this is the case we recommend preserving all editions of the raster representations for online display and zooming for 50 years.

After 50 years, raster base maps can be erased. The raster data of the more complete map can be thinned to the same interval as the 1:5,000 scale. Data of the same year should be preserved. When future technology is able to render legacy vector data online, only samples of the raster maps of this scale will need to be preserved. This might be the case before the 50-year time period elapses. Samples will serve as illustrations of the technology in use at the time of their production (requirement of some historians). Accompanying information about how the raster was generated should be preserved.

When there is thematic data at the 1:25,000 scale of a year where no raster map has been preserved the reference data in this scale can be generated from the preserved vector data.

5.2.3.5 Aerial photographs at a resolution of 50 cm per pixel

The only flight that senses all of Catalonia is divided into two resolutions of 25 cm and 50 cm per pixel. We recommend preserving permanently the first processable copy of the original image, which corresponds to the PAN.TIF and LR4.TIF files of the digital negative (see chapter 4.1.5). Due to the difference in quality, the RGB.TIF and CIR.TIF files cannot be considered redundant with the PAN.TIF and LR4.TIF and should be kept as well. The digital negative will be relevant for sciences such as biology or environmental science to have the most original data at the best resolution possible. Because RGB.TIF and CIR.TIF contain two redundant channels (red and green), we recommend investigating whether the channels can be merged to a single file containing all channels (red, green, blue and infrared) during a future storage hardware refreshment cycle.

Keeping the following details of the orthorectification process together with the digital negative would allow for maximum automation of regeneration of orthophotos:

- The algorithm to generate the RGBN master TIFF
- The look up table for the conversion of 12-bit to an 8-bit colour
- The georeference and orientation of the camera
- The digital terrain model used at the time
- The algorithm for orthorectification including the collinearity equation used
- The reference image for the radiometric correction.
- The seamlines for joining images
- Documentation of human intervention (type of actions taken, e.g., for quality control)

Although this information would not enable regeneration of the whole orthorectification process, the quality of the resulting orthophoto would still be better than what is currently available in other scales. Among the users' practice we could not determine tasks involving legacy geodata where remaining quality issues in regenerated orthophotos are problematic.

If a need for higher quality occurs in the future, remaining issues could still be solved through human intervention as it is done currently, or new algorithms might be developed. Historians that study the history of map making would have all the documentation about orthorectification available.

Users appreciate the immediate availability of orthophotos, which is why the generalized master copies at the 1:5,000 scale are available online. Due to the high popularity of aerial photographs in all user communities, we recommend keeping them available for 50 years. After this period, the time resolution of versions can be lowered to a five-year interval as long as requests for higher frequency can be satisfied with a regenerated orthophoto from the preserved negative. If this level of service cannot be reached, the archive will need to decide if the orthophotos should be kept for another 50 years or if alternative data sources cover the user needs. The digital negatives should also be described and linked in popular data catalogues so that users can discover them, especially when the lower resolution versions are not available any more.

Because of the national plan for aerial orthophotos[50] that decentralizes the production of this resolution, this data set has no redundant counterpart at the national level. The national data set of this scale is composed of the contributions of the regional authorities such as the ICGC. The National Geographic Institute’s copy is therefore equal to the one at the ICGC. Both institutions have the same open access strategy. Preservation responsibility can therefore be negotiated between these two institutions. If preservation will be decentralized as well, users will require visualization and download functionalities over the whole collection.

5.2.3.6 Aerial photographs at a resolution of 2.5 meters per pixel

This series is a generalization of the higher resolution orthophoto. We recommend keeping this series online for 50 years, to allow zooming out from the higher resolution orthophoto. During these 50 years, the use (online visualizations and downloads) should be carefully observed. This scale can be eliminated thereafter if

[50] Plan Nacional de Ortofotografía Aérea (PNOA). http://pnoa.ign.es/

- there were low request numbers for aerial photographs in the last few years and the generalization process can be performed on the fly for zooming, or
- other data sources at this scale can conveniently be integrated in the viewing service for zooming.

If neither of these conditions occurs, this resolution should be maintained at the same interval as the 1:5,000 scale. In any case the algorithm used for generalization should be documented.

5.2.3.7 Aerial photographs at a resolution of 10 cm per pixel

Several flights exist at this resolution – the coastal line (captured every year), urban areas (captured every four years) and the coastal zone (captured occasionally in 2010-2011). Even though these flights do not cover the whole territory of Catalonia they should be preserved because they represent the most detailed information available in the form of photographs. All digital negatives of this series should be preserved.

Aerial photographs of this resolution do not undergo the process of orthorectification explained in chapter 4.1.5. Orthophotos of these series can be visualized through the online service OrtoXpres. Because this service generates orthophotos on the fly, no product is saved on the producer side. The user has the option to download individual images.

Documentation of the way OrtoXpres rectifies the aerial photographs and the input needed for regeneration (approximate orientation of the camera at the time of capture and a digital terrain model) should be preserved.

5.2.3.8 Other aerial photographs

Socially important events of processes that occur suddenly, such as forest fires and floods, are subject to additional data capture. It is important to keep this data when it responds to general quality criteria. Social value and research value are guaranteed.

Aerial photographs of this kind are captured with another camera and do not undergo the standard process of orthorectification explained in chapter 4.1.5. Just as with the aerial photographs of resolution 10 cm per pixel, no orthorectified image is produced; only online visualizations through OrtoXpres are available. We recommend permanently preserving the digital negative that corresponds to a single RGBN.TIF[51] file of the original photograph. This data needs thematic keywords for researchers to access by subject (requirement of general public users, geographers and historians).

[51] This file contains the channels red, green, blue and near-infrared.

6 Conclusions

6.1 General goals achievement evaluation

This research was inspired by the requirement of the OAIS reference model to define designated communities. We wanted to define the user profiles for users of legacy geodata so that geodata archives can choose their designated communities based on the developed knowledge about those users (goal one). And we wanted to give users a voice in other decisions relevant to digital preservation of legacy geodata, in particular the development of appraisal criteria (goal two). We shall now analyse if these goals were reached.

Regarding goal one the first challenge was finding an appropriate method to define user profiles or designated communities. Lacking documented examples in the field of archival science, we made our own choice. We resorted to market research techniques to determine the user types and their characteristics, which we called modules. We found 16 types of legacy geodata users defined by a Delphi expert group, which was objective one. Objective two was reached with the definition of the modules (see chapter 5.1.1). Because of the similarity of module values and the similarity in functional needs, we showed that user types can be grouped into four user clusters interested in long-term legacy geodata: geographers, historians, urban planners and the general public. These four main user groups for legacy geodata are potential candidates for being chosen as designated communities of a long-term geospatial archive. Having extracted user characteristics of user types and clusters equates to having achieved goal one. Nevertheless, researching other user characteristics or deepening the knowledge about some of these characteristics such as the technological knowledge of the users in the group would bring archives even closer to alignment with the OAIS reference model. We will address this limitation of goal achievement in chapter 6.2.

As follows, we analyse the achievement of research goal two. In chapter 5.1.2.2 we have explained which user groups are relevant for a long term archive and why: the four clusters (objective four). Members of these relevant groups provided insight into their ways of working with legacy geodata in personal and focus group interviews. Even though, we propose to slightly change a part of the interview questions (see chapter 6.2 for the module 'acquisition interval') the largest part of the interview time was spent on open questions about participants' use of legacy geodata that we recommend to maintain. The transcripts of the interviews provided the rich description we intended to obtain in objective five.

After comparing existing appraisal criteria and guidelines for legacy geodata with user needs, little need for change of the criteria was detected. Our evaluations of the importance of some preservation arguments and a proposal of an amendment of the criteria spurred by user opinion are exposed in chapter 5.2.2.1 and summarized in chapter 6.4 achieving hereby objective six of this research. Conclusions from findings about additional user-informed legacy geodata preservation explained in chapters 5.2.2.2 and 5.2.2.3 can be adopted in the design of appraisal processes, service levels and components of a geospatial archival system (research goal two). In which context they can be adopted is analysed in chapter 6.5 about transferability. Finally, research objective seven was achieved with the proposition of specific preservation decisions for the main data sets of the ICGC in chapter 5.2.3. The decisions are founded on preferences of all four relevant user groups.

The following analysis of the modules and methods of this study explains the level of achievement of the research goals more in detail and determines if the objectives could have been achieved to a higher degree. We will make recommendations about the suitability of repeating the same modules, using the same methods to define user communities of other legacy data or applying the methods at other geodata archives. We will also analyse transferability to address the potential for the profiles to be reused by other geodata producers. Finally, we will suggest future research needs related to designated communities.

6.2 Evaluation of the modules

Due to lack of examples in the literature, we have chosen eight user characteristics based on different preservation guidelines that we thought relevant for assisting in making digital preservation decisions. The eight characteristics (modules) are: how the consumer interacts with the data (Interaction type), the data age range users are interested in (Time range), their technological knowledge (GIS knowledge), their preferred type of geodata product (Type of product), the scales they need (Scale), the ideal frequency of acquisition of versions and snapshots (Acquisition interval), the significance of the user group (Significance of the user group) and the preferred file formats (File format). Although the first research goal, to define user types and their characteristics, has been reached, not all modules proved to be useful for the purpose they were intended.

The interaction type and the type of product are related. Before using them for the selection of versions for the archive they must be put into context. It became obvious during interviewing that the type of product a person uses and his or her interaction with it is related to what is available and accessible for the area and era his or her research question is about. For example, where vector data is missing, the interaction type cannot be 'querying' even though users would like to do so. Additional influence comes from the user's technological knowledge. A user might have a research question for which querying would lead fastest to the answer, but he or she does not know how to do so.

The fact that the historic and research value increase with the age of the data, and the nonexistence of topographic vector databases older than 30 years, explains why current users do not explore legacy data in the sense of MacEachren. This should not hold us back from preserving vector data. When we noticed that the user had no choice in the type of product or interaction type, it was useful to ask about the hypothetical utility of alternative information sources as if they existed. The question about how users interact with data is easy to answer for them and seems easy to predict for a Delphi expert group.

Time range: The time range is very important because it determines the period data should stay in each archival stage, and it influences the point at which changes in acquisition intervals might occur. Delphi experts and user input coincide in this point for all four profiles, which is why we think both methods to determine the time frame are valid and either could be used to determine this module for other designated communities. The two modules, time range and acquisition interval, did successfully inform additional preservation decisions. Indeed, the two values could be directly applied to recommend ways of reducing the collection size.

GIS knowledge: The OAIS reference model requires an understanding of the designated community's knowledge base, to define the minimum representation information that must be maintained by an OAIS. The level of granularity to which we defined the 'GIS knowledge' of the user is not satisfactory for defining the representation information. We addressed technological software knowledge but did not collect sufficient information on familiarity with terminology nor go into detail with each mentioned software. For the definition of representation information, a user study dedicated exclusively to the renderability and understandability of preserved information to the designated community is needed. Necessary depth would be gained by facing users with specific chosen data objects and their terminology in a real access environment.

Scales: The information about the use of scales served to appraise data sets at the ICGC. This module has current validity because vector data are still captured at different scales. When multiscale databases are fully implemented, scales might be rendered automatically from a single data set. Thereafter raster maps will not be part of the production chain anymore and only the print version will have scales to choose from. The utility of this module is therefore limited in time. As with the type of interaction, we should put the scale in context with the available maps. Current users might use a specific scale, because the information they consult is only available at that scale.

Acquisition interval: The acquisition interval was the most difficult to determine. Few Delphi experts spoke out on the question and those who did seem to have ignored the distinction between vector and raster data that was made in the second round. When we asked about the ideal archive content, we hoped users would mention intervals of data they would like to encounter for their personal and work projects. This was not the case; participants were conditioned by the awareness of reality. Imagining the existence of historic data that was never created is imagining the impossible. Information about the ideal acquisition interval was not obtained in this way. Some users expressed their disappointment about the lack of information coming from certain decades, which indicates intervals that are too large. At the end of each interview, we asked about specific acquisition intervals but several users did not commit themselves to the question and expressed their disagreement with appraisal in general. Instead of asking participants to imagine the ideal archive for their work, perhaps we should have asked them to imagine a person with the same tasks as they have living 100 or 200 years from now and to reflect on what this future user would like us to keep of the current production. This might have given a more directly applicable answer.

Significance of the user group: The frequency of data use, the quantity of data used and the impact of the user group on the economy if legacy geodata were missing play a role in determining the significance of the user group. However, even though some users state they are either frequent or infrequent users, the data quantity they consult and download is difficult for them to judge because they do not have any reference to compare with. It is not possible to draw conclusions about the significance of the user group based on the interviews. Nevertheless, we received concrete quantified answers from the Delphi experts, who have an understanding of several user groups, which should be sufficient for making administrative decisions. The significance of the user group might have implications for decisions about the access infrastructure, the service level and the feasibility of preservation, which is why we recommend monitoring this module and using it in further user profiles. The economic significance of the user group might play a secondary role when the institution does not have to rely on income from the user, such as the ICGC.

File formats: Keeping data in various file formats causes redundancy. Therefore, an easy form of reducing the collection would seem to be selecting one out of many. Nevertheless, the file formats module could not be as directly applied to selection as the scale and the type of products, for several reasons. Currently, used file formats have to be kept to provide the different services. The user only interacts with the file formats that are distributed and not necessarily with the archived counterparts, which is why he or she cannot judge their utility. By the time data should pass to the semi-active or definitive archival stage and a file format is chosen for long-term preservation, the file formats available to users will likely differ from the ones the user interacts with currently. Therefore, current user preferences can only be taken into account as long as the file formats stay identical. The information on the file formats was not always easy to obtain. Sometimes file formats had to be guessed at from the explained interaction because users did not know this technical detail. Asking about the file formats a user prefers however was partly useful, because it brought up or confirmed the information about his or her technological knowledge. But because GIS knowledge is not sufficiently addressed with this type of user interview, we do not recommend using the file formats module for attempting to define other designated communities.

To summarize the reflection about the utility of the modules, we find that it is essential to have insight into the designated communities' knowledge base, ideal acquisition intervals and the data age the users are interested in. The significance of the user group can be helpful in some cases. While the acquisition interval and data age range module values could be directly applied to preservation decisions, all other information is much harder to apply because it is not quantified. Instead of separating information on the use of scales, the types of product and the interaction type it could be better collected by a rich description of the use cases. Finally, the module 'file formats' did not reveal useful. Questions about the file format might be asked in user interviews to bring up topics of technological knowledge.

6.3 Evaluation of the methods

Using the Delphi technique and the focus groups for the characterization of designated communities and preservation decisions is an original contribution of this research, because these techniques have never been used for this purpose before. Although focus of this thesis does not lay on the methods – the goal was not to propose a method – they can be partially evaluated because of the triangulation of different methods. In this chapter, we will assess the quality and feasibility of the Delphi method and the focus groups to make a recommendation about whether or not to repeat these methods in similar endeavours to define designated communities.

Delphi method:

The quality of a prediction is always difficult to assess unless the predicted events occur and can be measured. This makes it hard to talk about the quality of the module values predicted by the Delphi experts. Nevertheless, we can speak to the limits of the method. A Delphi study is a written method that relies on participants' understanding of the concepts and questions. Even though the most important terms were defined in the beginning, some answers suggested that misunderstandings still happened. Concepts such as what we meant by 'archiving' or 'raster data' should have been defined. Archiving has diverse meanings for the information science community and computer science people. The computer science term of a digital archive as a storage space or unit does not imply the standards and treatment an archivist would expect from a digital archive. Because people tend to be more familiar with computer science than with archival science, the long-term perspective gets lost when they think of a digital archive. Nevertheless, adding to the introduction of the survey form might have reduced the motivation to read it all, which would have entailed even lower agreement on the concepts.

The fact that the Delphi method has delivered results speaks for its effectivity, but cannot exclude other methods to be better. Because we could not find documented methods for defining designated community profiles, we had nothing to compare against the Delphi method, making it difficult to evaluate it as a method. The way we compared the Delphi predictions to the statements of the current users does not allow a formal assessment of the quality of the results either, for several reasons. First, the Delphi experts came from North America and several countries in Europe and the interviewed users came from Catalonia; for a formal comparison, the two groups should be representative of one another. Second, there is no known universe of legacy geodata users to contrast samples against for significance testing.

Focus groups:

There were several limitations to the focus group interviews. Even though we contacted a wide variety of users, we could not reach a representative of each user type that composed the final four clusters. This might have biased the general picture of the clusters. Classical focus groups have more participants; nevertheless, we did not encounter problems of group dynamics. In addition, as explained in the evaluation of the modules, the focus group results did not describe the knowledge base of the users in sufficient detail to determine the minimum needed representation information. Nevertheless, for the other modules the focus groups proved to be an effective method for collecting rich information on user characteristics.

Even though we would partly change the type of questions asked to current users, we think the two methods combine well because the Delphi predictions and user statements added context to each other and helped understand both. The differences in module values between the two methods could be explained by future development of technology and other environmental factors, variations in the way a module-related question was asked to the Delphi experts and the focus group participants or the sample of users interviewed (in the case of the urban planners).

The combination of the Delphi method with the focus group interviews produced a rich description of user characteristics and of what users do with geodata, and gave insight into future needs of similar users. Both methods have their advantages and disadvantages, which is why applying only one or the other would not have led to the same conclusions. The Delphi method seems sufficient for defining some user characteristics (first research goal), but thanks to direct user interaction we found that defining designated communities by profession is not so important, because people work in interdisciplinary groups. Regarding the second research goal (adapt/amend appraisal guidelines), the user interviews were important for bringing up the accessibility and usability issues. But they also showed user input to be biased in favour of current user needs and against potential future reuse, while the Delphi method was more neutral in this aspect.

Despite these limits, the Delphi method is considered sufficient for the definition of most module values because its comparison with the user input did not reveal major differences. A recommendation on which modules should be addressed or omitted in a Delphi study and which are better addressed through direct user interaction is given in chapter 6.5.

6.4 Lessons learned

The following major lessons learned should be taken into account when studying designated communities in the future.

The importance of defining designated communities by professional profiles decreases knowing that many tasks are done working in interdisciplinary groups. This is particularly true in academic fields but also for urban planners we interviewed, who were from the larger municipalities. The general public users are an exception here, as most of the interviewed members work alone when they search and handle legacy geodata. We can assume higher skills in designated communities that work in groups than what we would expect from individuals, because of the combined knowledge of group members. For example, if the professional vocabulary contained in the representation information is tested against a group of historians, it should be kept in mind that they probably work together with geographers when geodata is needed.

Knowing the characteristics of several user types is very important when making preservation decisions. Choosing a more general user group, such as the general public, as a designated community leads to adding more information to the representation information, which helps a wider community understand and use the data. Nevertheless, using only the characteristics of the general public for decisions about other preservation processes could be a threat to the needs of more specialized user groups of legacy data. Indeed, a high frequency of vector data acquisition is not needed for the general public, but it is a requirement of urban planners and many geographers. In consequence, we recommend taking into account all current user profiles to maximize reuse possibilities and therefore value for all user groups. This exemplifies the importance of having defined a complete set of user needs for each cluster. If the goal of another geodata archive is solely to define the representation information, it is enough to interview members of the chosen designated community. If the input will be used to make other preservation decisions, it is also important to inquire of other potential user groups and to address the time range and the acquisition interval users are interested in.

Regarding the goal to give the users a voice in appraisal; user interviews have confirmed our assumption that people are opposed to selection and want everything to be preserved. But we also showed that through qualitative analysis of use cases we can find indicators of where possible selection can be introduced. In this exercise, modules were of little utility. Rather than using the modules, we took into account all user input when revising the appraisal criteria for legacy geodata. In general, few changes had to be made to appraisal criteria: the importance of the feasibility criteria has decreased in the eyes of the users, while usability gained strength. Through the scales and file formats modules and the products of the production chain we evaluated the criterion ‘uniqueness’. The interviewed users did not attach importance to elimination of redundancies and substitutable data, quite the contrary; they take advantage of the offered choice. But when a data set is unique in the sense of being the only choice available, user attention is concentrated to that data set and the argument of uniqueness becomes highly important. Indeed, data sets that have no substitute are widely used. Knowledge about how users behave with scales assisted in interpretation and application of the criterion uniqueness on existing productions.

Most appraisal criteria were revealed to be related with the service level. Usability increases when metadata is complete and additional accessibility services exist. The organizational focus of the archive decreases in importance when catalogues are interconnected and data made findable this way. Potential future use can increase when data can be made more homogeneous (e.g., two different semantics are bridged). The creation and quality checking of metadata is a service that reflects on the criterion data quality. For its part, the service level influences the criterion feasibility. A high service level generates cost for the archive, which can reduce feasibility.

We should not deduce directly from the current users to the future designated community. Users expect and want all data to be preserved. Preserving data today helps all future users of legacy geodata, but does not necessarily help the current user today. Current users have no influence on what legacy geodata is available to them, and most users say that they can adapt to what is available. They refer to versions of the production chain, acquisition intervals, scales, file formats and geographic coverage. Current users do not lose time asking for the existence of lost or never-created historic data that would help resolving current tasks, they are primarily concerned with improving accessibility and usability. Pursuing improvement of accessibility and usability seems for them to be the more realistic endeavours. In the view of the current user it is not so important what we select, because this will only benefit future users, but more how we offer it, because this can affect him or her immediately. If we let appraisal decisions be guided completely by current user input we might increase service levels at the expense of data quantity, which would harm future use. Future use depends more on content existence than on service existence. When content is nonexistent, no use is made. When service is nonexistent, the user adapts data (e.g., creates georeferences or transforms file formats) or bypasses the fact that the service level does not exist (e.g., works in a group that can reach the desired data quality). Nevertheless, some service levels such as creating quality metadata, georeferencing and digitization benefit the present and the future. The art is to balance between services that benefit only current users and data quantity that future members of the designated community may need.

All recommendations made in this thesis are made from the perspective of the user and should favour reuse. They will not replace analysis of additional aspects that influence preservation decisions. Implementation of preservation processes might be guided by criteria that contradict acquisition intervals, time spans the data stays at each archival stage, appraisal criteria and service levels proposed in this thesis.

6.5 Transferability

Qualitative studies with purposeful sampling and an unknown universe like this research do not drive for generalization but for transferability of the results (Patton, 2015).[52] Transferability is given when the researcher shows that the results can be applied to similar individuals or institutions. First, we discuss transferability of the results. Then, we discuss whether the modules are useful for determining designated communities in other fields. Finally, we address feasibility of the methods used in this study in a professional context.

Regarding the first research goal, the characteristics of the different user groups have been mapped by study participants from Europe and North America and members of an international organization with western roots. The participating experts were assumed to have experience with users coming from the same environment and culture and that they had those user groups in mind when projecting into the future. Therefore, their answers predict characteristics of western legacy geodata users. The interviews were conducted with Catalonian users, while the Delphi study had a broader scope. Even though we have seen slight distinctions between the Delphi results and the user input from the interviews, we can consider the results as consistent because all differences could be explained by reasons other than the culture or geographic area. Our research method does not allow verifying to what degree user characteristics of other countries would vary from the Delphi answers. To know that, we would have to interview control groups outside Catalonia. Although we did not detect major differences in Catalonia, that does not mean that there are none in other regions. Hence, the 16 user types and 6 user clusters described by the Delphi experts and their module values cannot be generalized to be valid in all western countries. Nevertheless, taking into account differences that could occur due to culture such as the way archives are legislated and organized, the mapping culture, the penetration of technology or research trends and topics specific to the region, other western archives could adopt the user profiles as a basis for further decisions. Interested bodies might be archives, libraries, data

[52] 'A random and statistically representative sample permits confident generalization from a sample to a larger population.'

repositories and geodata producers or other institutions (with the intention of choosing designated communities for an OAIS.

To obtain similar detail as in this research, in each region the focus group interviews would have to be repeated.

The modules are product independent and can be reused; however, as explained in the evaluation of the modules, we would recommend using only the acquisition interval, time frame and significance of the user group modules. These three modules could be reused even in other disciplines not related to geodata. GIS knowledge should be addressed in a specific user study and not as module. To reuse this module in other disciplines user input collection should focus on the technologies of each field. In contrast, the module values depend on what is available to the user. The interviewed users in Catalonia mentioned and interacted with several data sources not produced by the ICGC. Therefore, the module values are transferable and can be applied accordingly at other geodata archives in Catalonia. Because the user profiles are easily separated into their individual modules, use cases needing input for only one or another of the characteristics is also imaginable. Because focus groups also informed the evaluation of the appraisal guidelines and other preservation decisions, those should be regarded as valid for Catalonia only.

In a professional context, these methods might seem too laborious to follow. But the process can be shortened if validity of the Delphi results outside Catalonia is assumed. Institutions in Europe and North America that are aware of the possible cultural differences can adapt the user profiles from the Delphi and the user-informed appraisal criteria. Nevertheless, they would still need to do a usability study for determining the representation information.

Even where the validity of the Delphi cannot be assumed, the process can still be simplified, because not all Delphi questions and user input was necessary. The Delphi study can be reduced to questions about the acquisition interval, the data age users are interested in and the significance of the user groups. Questions intending to predict change in user characteristics can be avoided, because they proved useless for the short period of time we asked about and would be even harder to predict for longer time frames. Subsequent user interviews should deliver a rich description of use cases and do not have to repeat the subjects already covered by the Delphi study, unless the latter did not produce valid answers. Finally, institutions in this category could do a usability study with legacy geodata for determining the representation information, which would go beyond the results of this thesis.

6.6 Further research

Given the users' desire that all geodata be preserved, we need research assessing the cost of preservation and the possibilities of alternative funding. Such research has been tackled already for digital preservation in general. Documentation of applications of the geoarchiving cost-benefit analysis guidance elaborated by the GeoMAPP project partners could build the basis for research in this sense.

The legacy geodata user profiles can provide a base for further research inquiring deeper into individual user types. This type of applied research is recommended for archives that choose to determine the representation information for their data sets. Other methods for defining designated communities should be tested and documented. Comparison of efficiency and effectivity of the methods could then lead to a well-argued recommendation for a method.

We have seen that dividing users by professional background is not always reliable. We suggest that further research on interaction type and reasons for legacy geodata use would help group users by other characteristics. This could be done using the input of the interview transcripts from this research study, without taking into account the classification of users made by the Delphi experts. Carving out types of questions users have to solve or their approaches to geographic problem solving might make predictions of what designated communities will use in the future more reliable.

7 Bibliography

If not otherwise mentioned, all URIs have been last accessed on June 14, 2017.

Abrams, B., Anderson, M., Cahill, C., Dove, L., Downs, R., Engle, E., … Ratcliffe, M. (2010). Framing a national strategy for the appraisal and selection of geospatial data. In *Digital Geospatial Appraisal Meeting Nov. 17-18, 2010* (Vol. 619). Library of Congress. Retrieved from http://www.digitalpreservation.gov/meetings/documents/othermeetings/Geo_appraisal_workshop_notes_final.pdf

Albani, M., Guarino, R., & Leone, R. (2011). *Long Term Data Preservation functional user requirements document LTDP / FURD*. https://doi.org/LTDP-GSEG-EOPG-RD-11-0004

Anguita, S., Montaner, C., Oller, J., & Roset, R. (2012). Digital preservation at the Institut Cartogràfic de Catalunya. *E-Perimetron*, *7*(2), 89–96. Retrieved from http://www.e-perimetron.org/vol_7_2/anguita et al.pdf

Aqdus, S. A., Hanson, W. S., & Drummond, J. (2012). The potential of hyperspectral and multi-spectral imagery to enhance archaeological cropmark detection: a comparative study. *Journal of Archaeological Science*, *39*(7), 1915–1924. https://doi.org/10.1016/j.jas.2012.01.034

Ariza López, F. J. (2012). Preservación de la información geográfica. *Revista Catalana de Geografia*, *vol XVII*(46). Retrieved from http://www.rcg.cat/articles.php?id=246

Auerbach, C. F., & Silverstein, L. B. (2003). *Qualitative data: an introduction to coding and analysis*. New York: New York University Press.

Baiocchi, V., & Lelo, K. (2010). Accuracy of 1908 high to medium scale cartography of Rome and its surroundings and related georeferencing problems. *Acta Geodaetica et Geophysica Hungarica*, *45*(1), 97–104. https://doi.org/10.1556/AGeod.45.2010.1.14

Ball, A. (2010). *Preservation and curation in institutional repositories* (DCC state of the art report). University of Edinburgh. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/reports/irpc-report-v1.3.pdf

Bayer, T. (2016). Advanced methods for the estimation of an unknown projection from a map. *GeoInformatica*, *20*(2), 241–284. https://doi.org/10.1007/s10707-015-0234-x

Berman, F., Kozbial, A., Mcdonald, R. H., & Schottlaender, B. E. C. (2008). The need to formalize trust relationships in digital repositories. *Educause Review*, *May/June*, 10–11. Retrieved from http://www.educause.edu/ir/library/pdf/ERM0835.pdf

Bernal, J. D. (1948). Preliminary analysis of pilot questionnaire on the use of scientific literature. In *Royal Society Scientific Information Conference, 21 June-2 July, 1948: Report and Papers Submitted* (pp. 589–637). London: Royal Society.

Beruti, V., Conway, E., Forcada, M. E., Giaretta, D., & Albani, M. (2010). ESA plans – a pathfinder for long term data preservation. In *iPRES 2010: proceedings of the 7th International Conference on Preservation of Digital Objects : September 19-24, 2010 - Vienna, Austria*. Retrieved from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beruti-76.pdf

Bethune, A., Lazorchak, B., & Nagy, Z. (2009). GeoMAPP: a Geospatial Multistate Archive and Preservation Partnership. *Journal of Map & Geography Libraries*, *6*(1), 45–56. https://doi.org/10.1080/15420350903432630

Biblarz, D., Tarin, M.-J., Vickery, J., & Bakker, T. (2001). *Guidelines for a collection development policy using the Conspectus Model*. Retrieved from http://www.ifla.org/files/assets/acquisition-collection-development/publications/gcdp-en.pdf

Bielecka, E., Leszczyńska, M., & J Halls, P. (2014). User perspective on geospatial data quality. Case study of the Polish Topographic Database. In *The 9th International Conference "Environmental Engineering", 22–23 May 2014, Vilnius, Lithuania* (pp. 1–6). Vilnius: Vilnius Gediminas Technical University Press Technika. https://doi.org/10.3846/enviro.2014.193

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boin, A. T., & Hunter, G. J. (2006). What communicates quality to the spatial data consumer? *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *34*(Part XXX), 140–147. Retrieved from https://pdfs.semanticscholar.org/877e/bd232160c6c8c6c99cd0c51d4e1a90ce4a6c.pdf

Bos, M., Gollin, H., Gerber, U., Leuthold, J., & Meyer, U. (2010). *Archiving of geodata: a joint preliminary study by swisstopo and the Swiss Federal Archives*. Retrieved from https://www.bar.admin.ch/dam/bar/en/dokumente/publikationen/vorstudie_zur_archivierungvongeodaten.pdf.download.pdf/preliminary_studyarchivingofgeodata.pdf

Bühler, J. (2005). A digital map library – electronic information in the map collection. *The Cartographic Journal*, *42*(3), 222–226. https://doi.org/10.1179/000870405X77156

California State Archives, The University of North Carolina at Chapel Hill, & California natural resources agency. (2010). *eLegacy: California's geospatial records: archival appraisal, accessioning, & preservation. Final narrative report*. Retrieved from http://salt.umd.edu/eLegacy/docs/FinalReport/eLegacy_final.pdf

Capdevila Subirana, J., Sánchez Maganto, A., Camacho Arranz, E., & Arístegui Cortijo, A. (2012). Opening up the cartographic heritage of the Spanish Geographical Institute by means of publishing standardized, Inspire compatible metadata. *LIBER Quarterly*, *22*(4), 345–357. Retrieved from https://www.liberquarterly.eu/articles/10.18352/lq.8049/

Caruso, G., Briguglio, L., Matthews, B., & Polsinelli, B. (2013). *D31.1 - Parameters for long term preservation and data sustainability models*. Retrieved May 9, 2015 from http://www.scidip-es.eu/scidip-es/deliverables.html

Catalonia. (1982). Ley 11/1982, de 8 de octubre, de creación del Institut Cartogràfic de Catalunya. *Diari Oficial de La Generalitat de Catalunya*, *268*, 2380.

Catalonia. (2001). Ley 10/2001, de 13 de julio, de archivos y documentos. *Boletín Oficial Del Estado BOE*, *206*(206), 32444–32453.

Catalonia. (2006). Ley 16/2005, de 27 de diciembre, de la información geográfica y del Instituto Cartográfico de Cataluña. *Diari Oficial de La Generalitat de Catalunya*, *4543*, 64–71.

Center for International Earth Science Information Network [CIESIN]. (2005). *Data model for managing and preserving geospatial electronic records, version 1.0*. New York: Columbia University. Retrieved from http://www.ciesin.columbia.edu/ger/DataModelV1_20050620.pdf

Center for Research Libraries. (2010). *Report on Portico Audit Findings*. Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/CRL Report on Portico Audit 2010.pdf

Center for Research Libraries. (2011). *Certification Report on the HathiTrust Digital Repository*. Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/CRL HathiTrust 2011.pdf

Center for Research Libraries. (2012). *Certification Report on Chronopolis*. Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/Chron_Report_2012_final_0.pdf

Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA, National Research Council, Division on Earth and Life Studies, & Board on Atmospheric Sciences and Climate. (2007). *Environmental data management at NOAA: archiving, stewardship, and access*. *Data Management*. Washington: The National Academies Press. Retrieved from http://www.nap.edu/catalog/12017.html

Committee on the Preservation of Geoscience Data and Collections, Committee on Earth Resources, National Research Council, & National Academy of Sciences. National Research Council. (2002). *Geoscience data and collections: national resources in peril*. The National Academy Press. Retrieved from http://www.nap.edu/catalog.php?record_id=10348

Consultative Committee for Space Data Systems [CCSDS]. (2011). *Audit and certification of trustworthy digital repositories: recommended practice CCSDS 652.0-M-1*. Retrieved from http://public.ccsds.org/publications/archive/652x0m1.pdf

Consultative Committee for Space Data Systems [CCSDS]. (2012). *Reference model for an Open Archival Information System (OAIS)* (version 2). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Conway, E., Pepler, S., Garland, W., Hooper, D., Marelli, F., Liberti, L., … Badiali, L. (2013). Ensuring the long term impact of earth science data through data curation and preservation. *Information Standards Quarterly*, *25*(3), 28. https://doi.org/10.3789/isqv25no3.2013.05

Craig, B. (2004). *Archival appraisal*. München: K.G. Saur.

Cruse, P., & Sandore, B. (2009). Introduction: the Library of Congress National Digital Information Infrastructure and Preservation Program. *Library Trends*, *57*(3), 301–314. https://doi.org/10.1353/lib.0.0055

De La Beaujardière, J. (2016). NOAA environmental data management. *Journal of Map & Geography Libraries*, *12*(1), 5–27. https://doi.org/10.1080/15420353.2015.1087446

Dent Goodman, V. (2011). *Qualitative research and the modern library*. Oxford: Chandos publishing.

Denzin, N. K. (1978). *The research act: a theoretical introduction to sociological methods* (2nd ed.). New York: McGraw-Hill.

Dingwall, G., Marciano, R., Moore, R., & Peters McLellan, E. (2005). From data to records: preserving the geographic information system of the City of Vancouver. *Archivaria*, *64*(Fall 2007), 181–198. Retrieved from http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13157/14408

Dobratz, S., & Schoger, A. (2007). Trustworthy digital long-term repositories: the nestor approach in the context of international developments. *Lecture Notes in Computer Science*, *4675*, 210–222. https://doi.org/10.1007/978-3-540-74851-9_18

Dorfey, B., Graf, S., Grau, B., Homberg, J., Huth, K., Keitel, C., … Wettmann, A. (2009). *Handreichung zur Archivierung elektronisch vorliegender Geodaten*. ARK AG ESys; ARK IT-Ausschuss. Retrieved from http://www.bundesarchiv.de/imperia/md/content/bundesarchiv_de/fachinformation/ark/handreichung_geodaten_20090928.pdf

Duerr, R. E., Weaver, R. L., & Kaminski, M. (2010). Data acceptance procedures and levels of service at the National Snow and Ice Data Center. In *2010 IEEE International Geoscience and Remote Sensing Symposium: proceedings* (pp. 2322–2325). Honolulu: IEEE. https://doi.org/10.1109/IGARSS.2010.5650356

Elzakker, C. P. J. M. van. (n.d.). *From map use research to usability research in geo-information processing*. Enschede: International Institute for Geo-Information Science and Earth Observation. Retrieved from http://cartography.tuwien.ac.at/wordpress/wp-content/uploads/2013/01/cartotalk-corne-van-elzakker.pdf

Elzakker, C. P. J. M. van. (2004). *The use of maps in the exploration of geographic data*. Utrecht University. Retrieved from https://www.itc.nl/library/Papers_2004/phd/vanelzakker.pdf

Elzakker, C. P. J. M. van, & van de Berg, W. P. E. (2010). Topographic base maps for physical planning maps: user research for generalization. In *Geospatial Data and Geovisualization: Environment, Security, and Society - Special Joint Symposium of ISPRS Technical Commission IV & AutoCarto 2010 in conjunction with ASPRS/CaGIS 2010 Special Conference*. Orlando: International Society for Photogrammetry and Remote Sensing. Retrieved from http://www.isprs.org/proceedings/XXXVIII/part4/files/van Elzakker.pdf

Erwin, T., & Sweetkind-Singer, J. (2009). The National Geospatial Digital Archive: a collaborative project to archive geospatial data. *Journal of Map & Geography Libraries*, *6*(1), 6–25. https://doi.org/10.1080/15420350903432440

Erwin, T., Sweetkind-Singer, J., & Larsgaard, M. L. (2009). The National Geospatial Digital Archives—collection development: lessons learned. *Library Trends*, *57*(3), 490–515. https://doi.org/10.1353/lib.0.0049

European Commission. High Level Expert Group on Scientific Data. (2010). *Riding the wave: how Europe can gain from the rising tide of scientific data. Final report*. European Union. Retrieved from http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707&usg=AFQjCNHsgaXvCSx8gz2mI0kMcoz5znEpzQ

European Union. (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. *Official Journal of the European Union*, (L 345), 90–96.

European Union. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union*, (L 108), 1–14.

Fildes, R., & Allen, P. G. (Eds.). (2011). *Forecasting: Vol.III: Judgemental methods and forecasting competitions*. Los Angeles: SAGE.

Frick, R., & Najar, C. (2009). *Historisierung, nachhaltige Verfügbarkeit und Archivierung von Geoinformation: eine Auslegeordnung*. Basel: SIK-GIS. Retrieved from http://www.sik-gis.ch/daten/SIK-GIS-Studie-Archivierung.pdf

Gantner, F., Waldvogel, B., Meile, R., & Laube, P. (2013). The basic formal ontology as a reference framework for modeling the evolution of administrative units. *Transactions in GIS*, *17*(2), 206–226. https://doi.org/10.1111/j.1467-9671.2012.01356.x

GeoConnections, & Hickling Arthurs Low Corporation. (2013). *Geospatial data preservation primer* (Canadian Geospatial Data Infrastructure, Information Product) (Vol. 36e). https://doi.org/doi:10.4095/296299

Geospatial Multistate Archive and Preservation Partnership [GeoMAPP]. (2011a). *Archival metadata elements for the preservation of geospatial datasets*. Retrieved from http://www.geomapp.net/docs/GIS_OAIS_Archival_Metadata_v1.0_final_20110921.pdf

Geospatial Multistate Archive and Preservation Partnership [GeoMAPP]. (2011b). *Best practices for archival processing for geospatial datasets*. Retrieved from http://www.geomapp.net/docs/GIS_Archival_Processing_Process_v1.0_final_20111102.pdf

Geospatial Multistate Archive and Preservation Partnership [GeoMAPP]. (2011c). *Geoarchiving comprehensive cost-benefit anaysis guidance*. Retrieved from http://www.geomapp.net/docs/03_Geoarchiving_Cost-Benefit_Analysis_Guidance_20111231.pdf

Geospatial Multistate Archive and Preservation Partnership [GeoMAPP], & Library of Congress. (2010). *GeoMAPP interim report: 2007-2009*. Retrieved from http://www.geomapp.com/docs/GeoMAPP_InterimReport_Final.pdf

Giaretta, D. (2011). *Advanced digital preservation*. Berlin: Springer.

Glaser, B. G., & Strauss, A. L. (1968). *The discovery of grounded theory: strategies for qualitative research*. London: Weidenfeld and Nicolson.

Gorman, G. E., & Clayton, P. (2005). *Qualitative research for the information professional: a practical handbook* (2nd ed.). London: Facet publishing.

Gregory, R. H., & Van Horn, R. L. (1960). *Automatic data-processing systems : principles and procedures*. San Francisco: Wadsworth Publishing Company.

Ground Segment Coordination Body. (2012). *Long term preservation of Earth Observation space data: European LTDP common guidelines*. Retrieved May 9, 2015 from http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue2.0.pdf

Ground Segment Coordination Body. (2015). *Long term preservation of Earth Observation space data: preservation guidelines*. Retrieved from https://earth.esa.int/documents/1656065/2265358/EO-Data-Preservation-Guidelines

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough?: An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Haklay, M., & Weber, P. (2008). OpenStreetMap: user-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Hank, C., Tibbo, H. R., & Barnes, H. (2007). Building from trust: using the RLG/NARA audit checklist for institutional repository planning and deployment. In *Archiving 2007: Arlington, Virginia, May 21 - 24, 2007; final program and proceedings* (pp. 62–66). Springfield: Society for Imaging Science and Technology. Retrieved from http://www.imaging.org/IST/store/epub.cfm?abstrid=34421

Harrie, L., Stigmar, H., & Djordjevic, M. (2015). Analytical estimation of map readability. *ISPRS International Journal of Geo-Information*, *4*(2), 418–446. https://doi.org/10.3390/ijgi4020418

Harris, R. (2001). Earth observation data archiving in the USA and Europe. *Space Policy*, *17*(1), 35–48. https://doi.org/10.1016/S0265-9646(00)00052-7

Henderson, J. (2006). Creating the GeoArchives - Maine Archives of Geographic Information: a brief report presented at the joint conference of NAGARA, SAA, and CoSA. In Society of American Archivists (Ed.), *Joint Annual Meeting of NAGARA, CoSA and SAA: held August 3-6, 2006*. St. Petersburg: Convention Recordings International.

Hoa Loranger, & Nielsen, J. (2006). *Prioritizing web usability*. Berkeley: New Riders Press.

Hoebelheinrich, N., & Banning, J. (2008). *An investigation into metadata for long-lived geospatial data formats*. Retrieved from http://www.ngda.org/reports/InvestigateGeoDataFinal_v2.pdf

Janée, G. (2008). Preserving geospatial data: the National Geospatial Digital Archive's approach. In *Archiving 2008: June 24 - 27, 2008, Bern, Switzerland; final program and proceedings* (pp. 25–29). Springfield: Society for Imaging Science and Technology.

Janée, G., Sweetkind-Singer, J., & Moore, T. (2009). *Final report of the National Geospatial Digital Archive (NGDA) and Federated Archive Cyberinfrastructure Testbed (FACIT) projects*. Retrieved from http://www.ngda.org/docs/ngda-final-report.pdf

JISC. (2007). *HEI records management: guidance on archival appraisal*. Retrieved from http://tools.jiscinfonet.ac.uk/bcs/archival-appraisal.pdf

Kanehira, T., Arakawa, Y., Yasumoto, K., & Wada, T. (2016). CURAP: CURating geo-related information on a mAP. In *IEEE International Conference on Consumer Electronics (ICCE)* (pp. 325–326). IEEE. https://doi.org/10.1109/ICCE.2016.7430631

Kärberg, T. (2014). Digital preservation and knowledge in the public archives: for whom? *Archives and Records*, *7962*(August), 1–18. https://doi.org/10.1080/23257962.2014.942606

Koch, T., & Denike, K. (2004). Medical mapping: the revolution in teaching—and using—maps for the analysis of medical issues. *Journal of Geography*, *103*(2), 76–85. https://doi.org/10.1080/00221340408978578

Kramers, E. R. (2008). Interaction with maps on the Internet – a user centred design approach for the Atlas of Canada. *The Cartographic Journal*, *45*(2), 98–107. https://doi.org/10.1179/174327708X305094

Krueger, R. A., & Casey, M. A. (2016). *Focus groups: a practical guide for applied research* (5th ed.). Los Angeles: SAGE.

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, *73*(5), 467–482. https://doi.org/10.1016/j.techfore.2005.09.002

Lazorchak, B., Nagy, Z., Goreham, D., Morris, S., Bethune, A., & Peters, M. (2008). Creating a business plan for the archival preservation of geospatial data. In *Archiving 2008: June 24 - 27, 2008, Bern, Switzerland; final program and proceedings* (p. 79). Springfield: Society for Imaging Science and Technology.

Leyk, S., Boesch, R., & Weibel, R. (2006). Saliency and semantic processing: extracting forest cover from historical topographic maps. *Pattern Recognition*, *39*(5), 953–968. https://doi.org/10.1016/j.patcog.2005.10.018

Litwin, L., & Rossa, M. (2011). *Geoinformation metadata in INSPIRE and SDI: understanding, editing, publishing*. New York: Springer.

Lohfink, A., Carnduff, T., Thomas, N., & Ware, M. (2007). An object-oriented approach to the representation of spatiotemporal geographic features. In *Proceedings of the Fifteenth International Symposium on Advances in Geographic Information Systems (ACM GIS 2007)*. Seattle, Washington.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems*. Hoboken: Wiley.

Lutz, M., & Kolas, D. (2007). Rule-based discovery in spatial data infrastructure. *Transactions in GIS*, *11*(3), 317–336. https://doi.org/10.1111/j.1467-9671.2007.01048.x

Ma, J., Smith, B. L., & Zhou, X. (2016). Personalized real-time traffic information provision: agent-based optimization model and solution framework. *Transportation Research Part C: Emerging Technologies*, *64*, 164–182. https://doi.org/10.1016/j.trc.2015.03.004

MacEachren, A. M., & Kraak, M. J. (1997). Exploratory cartographic visualisation: advancing the agenda. *Compurers & Geosciences*, *23*(4), 335–343. https://doi.org/10.1002/9780470979587.ch11

Malanowski, N., & Zweck, A. (2007). Bridging the gap between foresight and market research: integrating methods to assess the economic potential of nanotechnology. *Technological Forecasting and Social Change*, *74*(9), 1805–1822. https://doi.org/10.1016/j.techfore.2007.05.010

Mcgarva, G., Morris, S. P., & Janée, G. (2009). *Technology watch report: preserving geospatial data* (DPC Technology Watch Series). Digital Preservation Coalition. Retrieved from www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee

McKemmish, S. (1997). Yesterday, today and tomorrow: a continuum of responsibility. In *Proceedings of the 14th National Convention of the Records Management Association of Australia, 15-17 September 1997* (pp. 15–17). Perth: Records Management Association of Australia.

Millea, N. (2005). The LIBER Groupe des Carthothécaires map library usage survey, summer 2003: a mandate for change? *LIBER Quarterly*, *15*(1). Retrieved from urn:NBN:NL:UI:10-1-113403

Molch, K., Leone, R., Albani, M., & Mikusch, E. (2012). User needs and requirements impacting the long term preservation of earth observation data. In *2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2012): Munich, Germany, 22 - 27 July 2012* (Vol. d, pp. 7283–7285). Piscataway: IEEE. https://doi.org/10.1109/IGARSS.2012.6351980

Montaner, C. (2008). El proyecto de digitalización de la cartoteca del Institut Cartogràfic de Catalunya. *Revista Catalana de Geografia*, *vol XIII*(núm. 35). Retrieved from http://www.rcg.cat/articles.php?id=135

Montaner, C., & Capdevila Subirana, J. (2010). La información geográfica digital como patrimonio cartográfico. *Revista Catalana de Geografia*, *IV època*(núm. 40), 1–11.

Montaner, C., Pascual, V., & Roset, R. (2014). Geoportal IDE de mapas antiguos de Cataluña. *Revista Catalana de Geografia*, *vol XIX*(núm. 50). Retrieved from http://www.rcg.cat/articles.php?id=325

Moran, M. S., Hutchinson, B. S., Marsh, S. E., McClaran, M. P., & Olsson, A. D. (2009). Archiving and distributing three long-term interconnected geospatial data sets. *IEEE*

Transactions on Geoscience and Remote Sensing, *47*(1), 59–71. https://doi.org/10.1109/TGRS.2008.2002815

Morris, S. (2013). *Issues in the appraisal and selection of geospatial data: an NDSA report*. (National Digital Stewardship Alliance. Geospatial Content Team, Ed.). Retrieved from http://hdl.loc.gov/loc.gdc/lcpub.2013655112.1

Morris, S. P. (2010). *Appraisal and selection of geospatial data: prepared for Library of Congress*. Retrieved from http://www.digitalpreservation.gov/meetings/documents/othermeetings/AppraisalSelection_whitepaper_final.pdf

Morris, S. P., Nagy, Z., & Tuttle, J. (2010). *North Carolina Geospatial Data Archiving Project: final report*. Retrieved from http://www.digitalpreservation.gov/partners/ncgdap/high/ncgdap_final_report.pdf

Morris, S. P., & Tuttle, J. (2007). Curation and preservation of complex data: the North Carolina Geospatial Data Archiving Project. In *DigCCurr 2007: an international symposium in digital curation, April 18-20, 2007, Chapel Hill, NC, USA.* Chapel Hill. Retrieved from http://ils.unc.edu/digccurr2007/papers/tuttle_paper_4-3.pdf

Morris, S. P., Tuttle, J., & Essic, J. (2009). A partnership framework for geospatial data preservation in North Carolina. *Library Trends*, *57*(3), 516–540. https://doi.org/10.1353/lib.0.0050

Mountain, D., & MacFarlane, a. (2007). Geographic information retrieval in a mobile environment: evaluating the needs of mobile individuals. *Journal of Information Science*, *33*(5), 515–530. https://doi.org/10.1177/0165551506075333

NASA. (n.d.). *The Apollo 11 telemetry data recordings: a final report*. Retrieved from http://www.hq.nasa.gov/alsj/a11/Apollo_11_TV_Tapes_Report.pdf

National Research Council (U.S.). Committee on Research Priorities for the USGS Center of Excellence for Geospatial Information Science. (2007). *A research agenda for geographic information science at the United States Geological Survey*. Washington: National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12004

National Snow & Ice Data Center. (2006). *NSIDC DAAC Data Priority Workshop, January 11-12, 2006 Goddard Space Flight Center*. Retrieved from https://nsidc.org/sites/nsidc.org/files/files/data_workshop_report_2006.pdf

National Snow & Ice Data Center. (2013). *NSIDC DAAC product evaluation 2013: results report*. Retrieved from http://nsidc.org/sites/nsidc.org/files/files/NSIDC-DAAC-Product-Evaluation-2013-Results-Report.pdf

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, *70*(5), 56131. https://doi.org/10.1103/PhysRevE.70.056131

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Bridges between worlds: INTERCHI ’93 conference proceedings, Conference on Human Factors in Computing Systems, INTERACT ’93 and CHI ’93, Amsterdam, The*

Netherlands, 24 - 29 April 1993 (pp. 206–213). New York: ACM Press. https://doi.org/10.1145/169059.169166

Nogueras-Iso, J., Zarazaga-Soria, F. J., Lacasta, J., Béjar, R., & Muro-Medrano, P. R. (2004). Metadata standard interoperability: application in the geographic information domain. *Computers, Environment and Urban Systems*, *28*(6), 611–634. https://doi.org/10.1016/j.compenvurbsys.2003.12.004

North Carolina Geographic Information Coordinating Council. (2008). *GICC Archival and Long Term Access ad hoc Committee: final report*. Retrieved from https://ncit.s3.amazonaws.com/s3fs-public/documents/files/Archival-LongTermAccess-11-08-GICC.pdf

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), 15–29. https://doi.org/10.1016/j.im.2003.11.002

Palmer, C. L., Weber, N. M., & Cragin, M. H. (2011). Analytic potential of data. In *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11* (p. 425). New York: ACM Press. https://doi.org/10.1145/1998076.1998167

Patton, M. Q. (2015). *Qualitative research & evaluation methods: integrating theory and practice* (4th ed.). Los Angeles: SAGE.

Pérez, R. F., Pérez, Ó., Gutiérrez, L., Arnillas, S., Platania, V., Bemmelen, J. Van, & Sacramento, P. (2013). Towards long term data preservation of EO data in Europe and Canada with ESA's multi-mission PDGS. In *PV 2013 Ensuring Long Term Preservation and Adding Value to Scientific and Technical Data, ESA-ESRIN, Frascati, Italy*. Retrieved from https://www.researchgate.net/publication/275646159_Towards_Long_Term_Data_Preservation_of_EO_data_in_Europe_and_Canada_with_ESA%27s_Multi-Mission_PDGS

Pla, M., & Lleopart, A. (2010). Updating of vector databases at the Institut Cartografic de Catalunya. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Core Spatial Databases - Updating Maintenance and Services – from Theory to Practice* (Vol. XXXVIII, pp. 114–119). Haifa. Retrieved from http://www.isprs.org/proceedings/XXXVIII/4_8_2-W9/papers/final_136_ISPRS_Haifa_ICC_2010_Paper_20100127.pdf

Postma, T. J. B. M., Alers, J. C., Terpstra, S., & Zuurbier, A. (2007). Medical technology decisions in The Netherlands: how to solve the dilemma of technology foresight versus market research? *Technological Forecasting and Social Change*, *74*(9), 1823–1833. https://doi.org/10.1016/j.techfore.2007.05.011

Powell, A., Heaney, M., & Dempsey, L. (2000). RSLP collection description. *D-Lib Magazine*, *6*(9). https://doi.org/10.1045/september2000-powell

PREMIS Editorial Committee (Ed.). (2012). *PREMIS Data Dictionary for Preservation Metadata* (v. 2.2). Retrieved from http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

Projektteam Ellipse. (2012). *Projekt Ellipse. Konzeption der Archivierung von Geobasisdaten* *des Bundesrechts: Zwischenbericht*. Bern: Bundesamt für Landestopografie swisstopo; Schweizerisches Bundesarchiv BAR. Retrieved from https://www.bar.admin.ch/dam/bar/de/dokumente/publikationen/zwischenbericht_desprojektesellipse.pdf.download.pdf/zwischenbericht_desprojektesellipse.pdf

Projektteam Ellipse. (2013). *Projekt Ellipse. Konzeption der Archivierung von Geobasisdaten des Bundesrechts: Konzeptbericht*. Bern: Bundesamt für Landestopografie swisstopo; Schweizerisches Bundesarchiv BAR. Retrieved from https://www.bar.admin.ch/dam/bar/de/dokumente/konzepte_und_weisungen/konzeptbericht_ellipse.pdf.download.pdf/konzeptbericht_ellipse.pdf

Quisbert, H. (2008). Evaluation of a digital repository. In *Archiving 2008: June 24 - 27, 2008, Bern, Switzerland; final program and proceedings* (pp. 120–124). Bern: Society for Imaging Science and Technology. https://doi.org/urn:nbn:se:ltu:diva-29687

Reich, V., & Rosenthal, D. (2009). Distributed digital preservation: private LOCKSS networks as business, social, and technical frameworks. *Library Trends*, *57*(3), 461–475. https://doi.org/10.1353/lib.0.0047

Risbøl, O., Briese, C., Doneus, M., & Nesbakken, A. (2015). Monitoring cultural heritage by comparing DEMs derived from historical aerial photographs and airborne laser scanning. *Journal of Cultural Heritage*, *16*(2), 202–209. https://doi.org/10.1016/j.culher.2014.04.002

Rönsdorf, C., Mason, P., Holmes, J., Gerber, U., Streilein, A., Bos, M., … Stößel, W. (2013). *GI + 100: long term preservation of digital geographic information — 16 fundamental principles agreed by National Mapping Agencies and State Archives*. Retrieved from http://www.eurosdr.net/sites/default/files/images/inline/gi100_-_16_eurosdr_archiving_principles_v3.0.pdf

Rothenberg, J. (2012). Digital preservation in perspective: how far have we come and what's next [Video file]. Recordings of the speech at Future Perfect 2012, Wellington, NZ. Retrieved from https://www.youtube.com/watch?v=2Idbur1qR8I

Schuurman, N., & Leszczynski, A. (2006). Ontology-Based Metadata. *Transactions in GIS*, *10*(5), 709–726. https://doi.org/10.1111/j.1467-9671.2006.01024.x

Sears, G. (2001). *Geospatial data policy study: project report: executive summary*. Ottawa: KPMG Consulting. Retrieved from http://publications.gc.ca/collections/Collection/M4-1-2001E.pdf

Shaon, A., Rönsdorf, C., Gerber, U., Naumann, K., Mason, P., Woolf, A., … Samuelsson, G. (2011). Long-term sustainability of spatial data infrastructures: a metadata framework and principles of geo-archiving. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)*.

Shaon, A., & Woolf, A. (2011). Long-term preservation for spatial data infrastructures: a metadata framework and geo-portal implementation. *D-Lib Magazine*, *17*(9/10). https://doi.org/10.1045/september2011-shaon

Spain. (2010). Ley 14/2010, de 5 de julio, sobre las infraestructuras y los servicios de información geográfica en España. *Boletín Oficial Del Estado*, *163*(10707), 59628–59652.

Spain. (2011). Ley 23/2011, de 29 de julio, de depósito legal. *Boletín Oficial Del Estado*, *182*(13114), 86716–86727.

Steinhart, G. (2006). Libraries as distributors of geospatial data: data management policies as tools for managing partnerships. *Library Trends*, *55*(2), 264–284. https://doi.org/10.1353/lib.2006.0063

Stigmar, H., & Harrie, L. (2011). Evaluation of analytical measures of map legibility. *The Cartographic Journal*, *48*(1), 41–53. https://doi.org/10.1179/1743277410Y.0000000002

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: SAGE.

Suchan, T. A., & Brewer, C. A. (2000). Qualitative methods for research on mapmaking and map use. *The Professional Geographer*, *52*(1), 145–154. https://doi.org/10.1111/0033-0124.00212

Sweetkind, J., Larsgaard, M. L., & Erwin, T. (2006). Digital preservation of geospatial data. *Library Trends*, *55*(2), 304–314. https://doi.org/10.1353/lib.2006.0065

Tang, J., Song, Y., Miller, H. J., & Zhou, X. (2016). Estimating the most likely space-time paths, dwell times and path uncertainties from vehicle trajectory data: A time geographic method. *Transportation Research Part C: Emerging Technologies*, *66*, 176–194. https://doi.org/10.1016/j.trc.2015.08.014

The National Oceanic and Atmospheric Administration [NOAA]. (2007). *Environmental data management at NOAA: archiving, stewardship, and access: brief report*. *Sciences-New York*. The National Academies. Retrieved from http://dels.nas.edu/resources/static-assets/materials-based-on-reports/reports-in-brief/data_at_noaa_final.pdf

The National Oceanic and Atmospheric Administration [NOAA]. (2008). *NOAA procedure for scientific records appraisal and archive approval: guide for data managers*. Retrieved from https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA_Procedure_document_final.pdf

Tjalsma, H., & Rombouts, J. (2011). *Selection of research data: guidelines for appraising and selecting data*. Retrieved from https://dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANSselectionofresearchdata.pdf

Trinidad, A., Carrero, V., & Soriano, R. M. (2006). *Teoría fundamentada “Grounded Theory”: La construcción de la teoría a través del análisis interpretacional*. Madrid: Centro de Investigaciones Sociológicas.

United States of America. Executive Order 12906 of April 11, 1994. Coordinating geographic data acquisition and access: the National Spatial Data Infrastructure, 59, no. 71Federal Register 1–4 (1994).

Vickery, G. (2011). *Review of recent studies on PSI-re-use and related market developments*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/review-recent-studies-psi-reuse-and-related-market-developments

Vogels, M. F. A., de Jong, S. M., Sterk, G., & Addink, E. A. (2017). Agricultural cropland mapping using black-and-white aerial photography, Object-Based Image Analysis and Random Forests. *International Journal of Applied Earth Observation and Geoinformation*, *54*, 114–123. https://doi.org/10.1016/j.jag.2016.09.003

Walters, T. O., & Skinner, K. (2010). Economics, sustainability, and the cooperative model in digital preservation. *Library Hi Tech*, *28*(2), 259–272. https://doi.org/10.1108/07378831011047668

Weaver, R. L. S., Meier, W. M., & Duerr, R. M. (2008). Maintaining data records: practical decisions required for data set prioritization, preservation, and access. In *Proceedings of the 2008 IEEE International Geoscience and Remote Sensing Symposium* (Vol. III, pp. 617–619). https://doi.org/10.1109/IGARSS.2008.4779423

Wilson, T. (2008). The information user: past, present and future. *Journal of Information Science*, *34*(4), 457–464. https://doi.org/10.1177/0165551508091309

Worboys, M. (2005). Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, *19*(1), 1–28. https://doi.org/10.1080/13658810412331280167

Yang, W. dong, He, Y. hua, Sun, L. min, Lu, X., & Li, X. (2015). An optimal query strategy for protecting location privacy in location-based services. *Peer-to-Peer Networking and Applications*, 752–761. https://doi.org/10.1007/s12083-015-0328-0

Yuan, M. (2008). Adding time into geographic information system databases. In J. P. Wilson & A. S. Fotheringham (Eds.), *The Handbook of Geographic Information Science* (pp. 169–184). Oxford: Blackwell.

Annexes

1. Questionnaire sent to users of legacy maps
2. Delphi questions for group one, second round
3. Delphi questions for group two, second round
4. Semi-structured interview guide for the focus groups and personal interviews
5. Letter of consent for participants of the focus groups
6. Relation of users who participated in the questionnaire and focus group interviews

Annex 1: Questionnaire sent to users of legacy maps

Apreciado usuario de mapas históricos o antiguos, Barcelona, otoño de 2013

Soy doctoranda de la Universidad de Barcelona. En mi estudio intento averiguar cómo y por qué se usarán los mapas históricos en el futuro. El objetivo es saber qué funcionalidades y aspectos de los mapas de hoy se tienen que conservar a largo plazo. Quiero que los usuarios del futuro puedan utilizar los mapas de hoy en día de la misma manera que usted utiliza los históricos ahora, porque para el usuario futuro los mapas del 2013 serán antiguos. El objetivo es obtener criterios de selección y eliminación para productores de material cartográfico.

Conocer sus necesidades me ayudaría a estimar las de los usuarios futuros. Me podría ayudar explicándome por qué y cómo usa la cartografía histórica. ¿Puedo contar con su opinión?

Si acepta participar en la encuesta, le haré diez preguntas relacionadas con el objetivo de su uso de los mapas y de los aspectos y funcionalidades que más le importan, y le preguntaré su opinión sobre el futuro valor de los mapas actuales.

Hay varias formas de participar. Si ha recibido el cuestionario en papel, puede rellenarlo y enviarlo por correo postal a la dirección que figura abajo, o puede escribir a alocher@ub.edu para recibir una copia electrónica. Si lo ha recibido en formato electrónico, lo puede rellenar y enviar a la dirección alocher@ub.edu o también imprimirlo y enviarlo a la dirección postal.

Gracias por su atención,

Anita Locher alocher@ub.edu
Facultad de Biblioteconomía y Documentación / Universidad de Barcelona
C/ Melchior de Palau 140
08014 Barcelona

****

Benvolgut usuari de mapes històrics o antics, Barcelona, tardor 2013

Sóc doctoranda de la Universitat de Barcelona. En el meu estudi intento esbrinar com i per què s'usaran els mapes històrics en el futur. L'objectiu és saber què funcionalitats i aspectes dels mapes d'avui s'han de conservar a llarg termini. Vull que els usuaris del futur puguin utilitzar els mapes d'avui dia de la mateixa manera que vostè utilitza els històrics ara, perquè per a l'usuari futur els mapes del 2013 seran antics. L'objectiu és obtenir criteris de selecció i eliminació per a productors de material cartogràfic.
Conèixer les seves necessitats m'ajudaria a estimar les dels usuaris futurs. Em podria ajudar explicant-me per què i com usa la cartografia històrica. Puc comptar amb la seva opinió?

Si accepta participar en l'enquesta, li faré vuit preguntes relacionades amb l'objectiu del seu ús dels mapes i dels aspectes i funcionalitats que més li importen, i li preguntaré la seva opinió sobre el futur valor dels mapes actuals.

Hi ha diverses formes de participar. Si ha rebut el qüestionari en paper, pot emplenar-ho i enviar-ho per correu postal a l'adreça que figura a baix, o pot escriure a alocher@ub.edu per rebre una còpia electrònica. Si ho ha rebut en format electrònic, ho pot emplenar i enviar a l'adreça alocher@ub.edu o també imprimir-ho i enviar-ho a l'adreça postal.

Gràcies per la seva atenció,

Anita Locher alocher@ub.edu
Facultat de Biblioteconomia i Documentació / Universitat de Barcelona
C/ Melchior de Palau 140
08014 Barcelona

****

Cuestionario dirigido a usuarios de mapas históricos o antiguos (mapas que no sean actuales).

****

Primera parte:

Le pedimos algo de información sobre usted.

1) **¿Qué es su experiencia?** Si hizo estudios superiores, Indica el grado o la licenciatura que cursó o está cursando. Si no tiene carrera académica, indica con qué grupo profesional se identifica más.

□ Geografía

□ Historia

□ Geologia

□ Arquitectura

□ Arqueología

□ Otros. Por favor, especifíca: ____________________

****

Segunda parte:

Le pedimos información sobre su **última consulta de mapas antiguos** de la cartoteca. Si está estudiando mapas con regularidad le rogamos que respondan al cuestionario pensando en los objetivos de su último estudio.

2) ¿Con qué objetivo general estudia mapas antiguos o históricos?

Marca todas las opciones que corresponden.

□ Estudios históricos geográficos

□ Estudios históricos económicos

□ Estudios históricos sociales

□ Estudios artísticos / de diseño

□ Arqueología

□ Genealogía

□ Obligación legal, litigio etc.

□ Otros. Por favor, especifíca: ____________________

3) ¿Cuántos años tienen los mapas históricos o antiguos que usa?

Marca todas las opciones que corresponden.

□ 1-10 años

□ 10-50 años

□ 50-100 años

□ más de 100 años

4) ¿Tiene que comparar varios mapas para alcanzar su objetivo?

□ No □ Sí. Por favor, especifique lo que compara:

Por ejemplo: la misma región en varios años o varios mapas de la misma época o serie o el diseño por varios autores o la evolución de la toponímia etc.

5) Indique la importancia que tienen los siguientes aspectos del mapa para alcanzar su objetivo

	no aplica o, no sé	no importa	Es algo importante	Es muy importante
Coloración y diseño del mapa	○	○	○	○
La simbología y leyenda usada en el mapa	○	○	○	○
La precisión y exactitud del mapa	○	○	○	○
La proyección usada	○	○	○	○
El aspecto físico del mapa	○	○	○	○
La calidad del mapa = Carga o falta de objetos	○	○	○	○
El año de edición o de creación	○	○	○	○
La escala del mapa	○	○	○	○
La homogeneidad de los mapas que comparo	○	○	○	○
El idioma del mapa	○	○	○	○
La orientación del mapa	○	○	○	○
Otros, por favor especifica:	○	○	○	○

6) Una vez tiene un mapa relevante en sus manos, ¿qué dificultades se encuentra en usarlo o interpretarlo? O ¿qué funcionalidades le gustaría que tuviese el mapa?

7) Los mapas vectoriales digitales combinados con un programa adecuado permiten calcular distancias y superficies. También podría buscar objetos representados de una categoría concreta, de un cierto tamaño o a una cierta distancia de otro objeto. **¿Le ayudaría para su objetivo si el mapa que está utilizando fuese vectorial digital?**

Por favor marca la casilla que corresponde:

○ Sí ○ No ○ No lo sé

8) ¿En el marco de su estudio, ha vuelto o volverá a crear un mapa para representar la información que encuentra en los mapas históricos?

○ Sí ○ No ○ No lo sé

****

Tercera parte:

Le pedimos que haga una estimación del futuro.

9) Hoy en día los mapas modernos se crean en digital y se pueden consultar a través de diferentes herramientas como Google maps, el GPS del coche o aplicaciones específicas como el Vissir del Institut Cartogràfic de Catalunya. Si conseguimos preservarlos los usuarios del futuro los podrán consultar. **¿Cree que a un usuario, que vive en el año 2063 y que tenga el mismo objetivo que usted, le pueden ser útiles los mapas del 2013?** Por favor marca la casilla que corresponde:

○ Creo que sí. ○ Creo que no.

Por favor, explique ¿por qué?

10) Cree que en 50 años todavía habrá usuarios con los mismos objetivos que usted? Por favor marca la casilla que corresponde:

○ Creo que sí. ○ Creo que no.

Por favor, explique ¿por qué?

Si quiere puede explicar un poco más lo que hace o aportar otros comentarios:

Por favor envíe el cuestionario a alocher@ub.edu o mándelo por correo postal a la dirección que aparece en la primera página **antes del 18 de noviembre de 2013**. ¡Muchas gracias!

Annex 2: Online Delphi survey (group one, second round)

2nd round of the Delphi study on use of superseded geodata

Welcome to the second round of questions for the Delphi study on digital maps and file formats.

The goal of this second round is to give you feedback on the answers of the other experts and to narrow the deviation in the group answers in order to reach a more precise estimation.

I would also like to emphasize that for the sake of this study I use the term geodata only for data that is used to create maps. I do not include satellite data or seismic or meteorological data.

Please remember the following definitions used in the study:

Vector data base : the vector content of a data base or GIS.

Vector map : the vector data in a processed form, represented as a map and in a sharable vector file format. An ESRI shapefile is a vector map.

Geospatial data = geodata: vector and raster data needed for creating maps. This includes remote sensed data, elevation models, data bases, vector and raster maps but excludes thematic tabular data for the sake of this study.

Version : every time a map layer or data base content is updated a new version is created. In some cases the new version overwrites the old information but modern GIS can associate a date to each update so that a sequence of versions is created and the historic development can be reconstructed. In remote sensing a new version is created when the same area has been photographed (or sensed) at the same scale for a second or subsequent time. For edited maps the new version is the new updated edition.

Superseded data: all versions that are not the most current. I prefer to talk about superseded data rather than historic, because there is a certain expectation for historic data to be antique. This study covers all ages of superseded data: the data that just became superseded one week ago as well as the data that is more than 100 years old.

Snapshot: a snapshot is a copy that represents the state of a data base at a certain moment in time.

Time series: several versions of the same data set form a time series. Every version is created at a different moment in time, usually at regular intervals, therefore the name.

There are 59 questions in this survey

The user groups

1 [present_users]

These are the user groups we used in first round of this study:

- Historians
- Geographers
- General public
- Archaeologists
- Policy makers (Government)
- Architects and geo-related engineers
- Cultural and arts documentalists (i.e. filmmakers or museum staff)
- Lawyers
- Conservation agents of the non-natural environment (I.e. ancient building inspectors)
- Environmentalists of the natural environment
- Emergency response planners
- Undergraduate teachers and students
- Institutions that maintain and updata geodata (i.e. national survey institutions)

The following additional groups were mentioned in the first round: social scientists, seismologists, geologists, tourists, journalists, (urban) planners, defence/military, ONGs (human development), statisticians, professions affected directly by the weather (fishermen etc.), outdoor athletes, utility companies, retailers, banks, insurances, construction companies.

Journalists, ONG's and statisticians will join the new group of "Social scientists".

Seismologists and geologists will be in the new user group of "Geophysicists".

Tourists and outdoor athletes are included in the "General public".

Planners and urban planners will join the group of "Policy makers".

Defence and military will join the group of "Emergency respondent planning teams".

Users of weather related data will be excluded, as this study concentrates on data that is used for map creation.

Retailers, banks, utility companies and insurance companies will form the new group of "Commercial users".

Regarding App-users; those who create them or publish geodata through them will be regarded as "Commercial users" those who download and use the apps will pertain either to "General public" or to one of the other professions.

Finally, construction companies are to be included in the group of "Architects and geo-related engineers".

This results in the following new user groups:

- Geophysicists (including seismologists, geologists)
- Social Scientists (including economists, journalists, statisticians, people involved in development)
- Commercial users (including banks, utility companies, retailers etc.)

Please specify below how familiar you are with the new user groups. If you have not participated in the first round of this study, please, specify for all user groups.

2 [familiar_with]

How familiar are you with the following user groups?

Please choose the appropriate response for each item:

	I have no experience with this user group	I have some experience with this user group	I have a lot of experience with this user group
Geophysicists (including seismologists and geologists)	○	○	○
Social scientists (statisticians, journalists, human development etc.)	○	○	○
Commercial users (banks, utility companies, insurances etc.)	○	○	○
General public (tourists, outdoor sports etc.)	○	○	○
Archeologists	○	○	○
Historians	○	○	○
Geographers	○	○	○
Lawyers	○	○	○
Policy makers (government) and planers	○	○	○
Culture and art documentalists (i.e. filmmakers and museum staff)	○	○	○
Emergency respondent planning members and military	○	○	○
Conservation agents (non-natural environment. I.e. ancient building inspectors)	○	○	○
Environmentalists of natural environment (public and private)	○	○	○

	I have no experience with this user group	I have some experience with this user group	I have a lot of experience with this user group
Architects, engineers and construction	○	○	○
Institutions that update and maintain geodata	○	○	○
Undergraduate teachers and students	○	○	○

You could be familiar with those user groups because they are part of your institutions clients, or just because you have some knowledge about their needs.

What users do with data

3 [type_of_use]

The question was: What do these user groups mainly do with superseded data?

The following are the results from the first round:

The majority said that the following user groups mainly consult or look at the superseded data:

- General public
- Historians
- Lawyers
- Policy makers
- Culture and art documentalists
- Emergency respondent planning members
- Conservation agents
- Environmentalists (public and private)

The majority said that the following user groups mainly change and manipulate the superseded data:

- Arqueologists
- Geographers
- Architects and engineers
- Institutions that update and maintain geodata
- Undergraduate teachers and students

Please determine the main use for the new user groups. If desired, change your guess for the other user groups.

Please choose the appropriate response for each item:

	Consult/visualize the data (look at the data)	Change/manipulate the data

	Consult/visualize the data (look at the data)	Change/manipulate the data
Geophysicists	○	○
Social Scientists	○	○
Commercial users	○	○
General public	○	○
Archaeologists	○	○
Historians	○	○
Geographers	○	○
Lawyers	○	○
Policy makers	○	○
Culture and art documentalists (i.e. filmmakers, museum staff)	○	○
Emergency respondent planning members	○	○
Conservation agents	○	○
Environmentalists (public and private)	○	○
Architects and engineers	○	○
Institutions that update and maintain geodata	○	○
Undergraduate teachers and students	○	○

4 [My_knowledge2]

Regarding the previous question, how do you feel?

Please choose **only one** of the following:

- ○not confident about your answer
- ○little confidence about your answer
- ○quite confident about your answer
- ○confident about your answer

5 [not_confident]

If you really do not feel confident at all consider passing this question to a collegue or letting the "no answer" option selected.

Only answer this question if the following conditions are met:
° ((**My_knowledge2.NAOK** == "A1"))

What users do with data

6 [access_evolves]

How do you feel about the following statement:

"When more superseded data is made available online, access percentage by "general public" will rise compared to the other user groups."

Please choose **only one** of the following:

- ○I strongly agree
- ○I somehow agree
- ○Neither agree nor disagree
- ○I somehow disagree
- ○I strongly disagree

7 [access_percentage]

The question was: The sum of the access' of all users is 100%. Estimate the percentage of access of each user group.

If a professional accesses data in his free time for personal purpose, this type of access should be counted with the "general public".

The following table shows the arithmetic mean of participants answer in the first round:

User groups	Percentage of access
General public	21
Archaeologists	13
Historians	14
Geographers	11
Lawyers	5
Policy makers (government)	9
Culture and art documentalists (i.e. filmmakers and museum staff)	6
Emergency respondent planning members	4
Conservation agents (non-natural environment. I.e. ancient building inspectors)	4
Environmentalists of natural environment (public and private)	5
Architects and engineers	7
Institutions that update and maintain geodata	9
Undergraduate teachers and students	8
Total	116

The above table shows a total of access which is more than 100%. This needs to be corrected in the second estimation. Please take into account the new user groups and make a new estimation on the percentage of access of each user group..

Please write your answer(s) here:

- Geophysicists (including seismologists, geologists)
- Social scientists (including economists, journalists, statisticians, ONGs etc.)
- Commercial users (including banks, utility companies, retailers etc.)
- General public
- Archaeologists

- Historians
- Geographers
- Lawyers
- Policy makers (government)
- Culture and art documentalists (i.e. filmmakers and museum staff)
- Emergency respondent planning members
- Conservation agents (non-natural environment. I.e. ancient building inspectors)
- Environmentalists of natural environment (public and private)
- Architects and engineers
- Institutions that update and maintain geodata
- Undergraduate teachers and students

For example, estimate that 40% of access comes from historians, 15% from geographers, 10% from architects etc.

[Question 8 and 9 are equal to question 3 and 4]

The data age range

10 [age_range_evolves]

How do you feel about the following statement?

"The proportion of available recent data will increase in the next 10 years, while that of older data will decrease."

Please choose **only one** of the following:

- ○I strongly agree
- ○I somehow agree
- ○Neither agree nor disagree
- ○I somehow disagree
- ○I strongly disagree

11 [statement_changerate]

How do you feel about the following statement?

"There will be more interest in relatively recent data as the update rate in some data increases."

Please choose **only one** of the following:

- ○I strongly agree
- ○I somehow agree
- ○Neither agree nor disagree

- ○I somehow disagree
- ○I strongly disagree

12 [influence_ofavailabl]

How do you feel about the following statement?

"The availability of data influences the data age range in which a user group is interested."

Please choose **only one** of the following:

- ○I strongly agree
- ○I somehow agree
- ○Neither agree nor disagree
- ○I somehow disagree
- ○I strongly disagree

13 [map_age_range]

The question was: Estimate the map-age range in which the following user groups are most interested:

The unit is years. "999" was used for "as old as possible".

The following table shows the result from the first round:

User groups	Median of the minimum data age estimated	Median of the maximum data age estimated
General public	1	999
Archaeologists	1	999
Historians	10	999
Geographers	1	70
Lawyers	1	80
Policy makers (government)	1	75
Culture and art documentalists (i.e. filmmakers and museum staff)	1	999
Emergency respondent planning members	1	20
Conservation agents (non-natural environment. I.e. ancient building inspectors)	1	999
Environmentalists of natural environment (public and private)	1	200
Architects and engineers	1	100
Institutions that update and maintain geodata	1	200
Undergraduate teachers and students	1	150

Please estimate the maximum and minimum data age in which the new user groups are interested. If desired, change the estimation that you made for the previous user groups.

	Minimum data age this user group is interested in	Maximum data age this user group is interested in (use 999 for "as old as possible")
Geophysicists		
Social scientists		
Commercial users		
General public		
Archaeologists		
Historians		
Geographers		
Lawyers		
Policy makers (government)		
Culture and art documentalists		
Emergency respondent planning members		
Conservation agents (non-natural environment)		
Environmentalists of natural environment (public and private)		
Architects, engineers and construction		
Institutions that update and maintain geodata		
Undergraduate teachers and students		

Please indicate the age range in years **from 1** (one year old data or every update) **to 999** (for "as old as possible")

[Question 14 and 15 are equal to question 3 and 4]

Requested data size

16 [amounts_of_data]

The question was: Please select the user groups that might request the biggest amount of data (in data size not in land coverage).

The answers from the first round identified the following user groups as those that request the biggest amount of data:

- Geographers
- Institutions that update and maintain geodata
- General public
- Undergraduate teachers and students

Taking into account the three new user groups and focusing on superseded data, please select the user groups that might request the biggest amount of superseded data.

Please select at least 3 answers

Please choose **all** that apply:

- □Geophysicists
- □Social scientists

- □Commercial users
- □General public
- □Archaeologists
- □Historians
- □Geographers
- □Lawyers
- □Policy makers and planners
- □Culture and art documentalists
- □Emergency respondent planning members and military
- □Conservation agents
- □Environmentalists (public and private)
- □Architects, engineers and construction
- □Institutions that update and maintain geodata
- □Undergraduate teachers and students

[Question 17 and 18 are equal to question 3 and 4]

Types of data sets

19 [GIS_mainstream]

How many years might it take for 50% of the general public to be comfortable with inquiring vector geodata.

Please estimate taking into account the developments in technology, usability, education ect.

Please choose **only one** of the following:

- ○0-5 years
- ○5-10 years
- ○10-20 years
- ○more than 20 years
- ○never

[Question 20 and 21 are equal to question 3 and 4]

22 [GIS_web_enabled]

How many years might it take for 80% of all implemented GIS services to be web-enabled?

Please choose **only one** of the following:

- ○0-5 years
- ○5-10 years
- ○10-20 years

- ○more than 20 years
- ○never

[Question 23 and 24 are equal to question 3 and 4]

25 [left_shift]

How do you feel about the following statement:

"In the next 10 years, usage patterns will shift slightly towards the left (less usage of end products, more usage of intermediate products)."

Please choose **only one** of the following:

- ○I strongly agree
- ○I somehow agree
- ○Neither agree nor disagree
- ○I somehow disagree
- ○I strongly disagree

26 [type_of_data]

The question was: Imagine that all data is available without any restrictions or costs for the user. Which are the data sets that these user groups are most interested in?

The following table shows the results from the first round:

User groups	Taking into account all answers without weighting them	Weighting the answers taking into account participants answers to the question "how familiar they are with the user group" and "How confident they feel about their answer"
General public	Orthocorrected photography of earth	Raster map / Orthocorrected photography of earth (1)
Archaeologists	Slight tendency for Orthocorrected photography of earth	No clear tendency towards one data set (2)
Historians	Raster map	Raster map
Lawyers	Slight tendency for Raster map	Not enough answers (3)
Policy makers	No clear tendency towards one data set	Raster or vector map
Culture and art documentalists	Raster map	Not enough answers
Emergency respondent planning members	Not enough answers	Not enough answers
Conservation agents	GIS data base with vector data / Raster map	Not enough answers
Geographers	GIS data base with vector data	GIS data base with vector data
Environmentalists (public and private)	Not enough answers	Not enough answers
Architects and engineers	No clear tendency towards one data	GIS data base with vector data

	set	
Institutions that update and maintain geodata	GIS data base with vector data	GIS data base with vector data
Undergraduate teachers and students	Raster map	Raster map

(1) If two data sets are mentioned it means that both got at least 3 votes with a maximum of 1 vote apart.

(2) "No clear tendency towards one data set" means that no data set has reached the majority.

(3) "Not enough answers" means that there were 3 or less answers for that data set.

Please guess in which superseded data set the new user groups are most interested. If desired, change the estimation you made for the previous user groups.

Please choose the appropriate response for each item:

	Raw aerial photography	Raw lidar data	Digital elevation models (DTM and DSM)	Orthocorrected photography of earth	GIS data base with vector data	Vector map (shareable file)	Raster map
Geophysicists	○	○	○	○	○	○	○
Social scientists	○	○	○	○	○	○	○
Commercial users	○	○	○	○	○	○	○
General public	○	○	○	○	○	○	○
Archaeologists	○	○	○	○	○	○	○
Historians	○	○	○	○	○	○	○
Geographers	○	○	○	○	○	○	○
Lawyers	○	○	○	○	○	○	○
Policy makers (government) and planners	○	○	○	○	○	○	○
Culture and art documentalists	○	○	○	○	○	○	○
Emergency respondent planning members and military	○	○	○	○	○	○	○
Conservation agents (non-natural environment)	○	○	○	○	○	○	○
Environmentalists of natural environment (public and private)	○	○	○	○	○	○	○
Architects, engineers and construction	○	○	○	○	○	○	○
Institutions that update and maintain geodata	○	○	○	○	○	○	○
Undergraduate teachers and students	○	○	○	○	○	○	○

[Question 27 and 28 are equal to question 3 and 4]

Users with GIS knowledge

29 [GIS_knowledge]

The question was: What percentage of the potential users in the following groups have knowledge in GIS technology? Please estimate the % for each user group.

The following table shows the median of the group answers from the first round:

User groups	Percentage that has GIS knowledge	Standard deviation
General public	5	14,1
Archeologists	25	13,5
Historians	5	5,7
Geographers	90	27
Lawyers	1	4,3
Policy makers (government)	5	12,5
Culture and art documentalists (i.e. filmmakers and museum staff)	1	11,3
Emergency respondent planning members	30	32,4
Conservation agents (non-natural environment. I.e. ancient building inspectors)	6	27,4
Environmentalists of natural environment (public and private)	20	20,3
Architects and engineers	47	27,6
Institutions that update and maintain geodata	95	4,4
Undergraduate teachers and students	5	23,2

Please estimate the percentage for the new user groups. If desired, change the estimation you made for the previous user groups.

Please write your answer(s) here:

- Geophysicists
- Social scientists
- Commercial users
- General public
- Archaeologists
- Historians
- Geographers
- Lawyers
- Policy makers (government) and planners
- Culture and art documentalists (i.e. filmmakers and museum staff)
- Emergency respondent planning members and military

- Conservation agents (non-natural environment. I.e. ancient building inspectors)
- Environmentalists of natural environment (public and private)
- Architects, engineers and construction
- Institutions that update and maintain geodata
- Undergraduate teachers and students

[Question 30 and 31 are equal to question 3 and 4]

Negative economic impact

32 [economic_impact]

The question was: Which user groups would have the biggest negative economic impact on society if no superseded geodata were available. Please rank from the biggest (on the top) to the smallest (at the bottom) negative economic impact.

The following table shows the weighted group answer from the first round

Rank	User group
1	Institutions that update and maintain geodata
2	Geographers
3	General public
4	Environmentalists of natural environment (public and private)
5	Architects and engineers
6	Policy makers (government)
7	Emergency respondent planning members
8	Historians
9	Undergraduate teachers and students
10	Culture and art documentalists
11	Lawyers
12	Conservation agents (non-natural environment)
13	Archaeologists

Please compare the new user groups to the previous. Imagine the new user groups had no access to superseded geodata. Would they generate a similar negative economic impact on society as one of the other user groups? Please select the one that has the most similar impact:

33 [impact_geophysicists]If geophysicists had no access to superseded geodata this would create a similar negative impact on society as...

Please choose **only one** of the following:

- ○General public
- ○Archaeologists

- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers (government) and planners
- ○Culture and art documentalists (i.e. filmmakers and museum staff)
- ○Emergency respondent planning members and military
- ○Conservation agents (non-natural environment)
- ○Environmentalists of natural environment (public and private)
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

34 [impact_social_scienc]

If social scientists had no access to superseded geodata this would create a similar negative impact on society as...

Please choose **only one** of the following:

- ○General public
- ○Archaeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers (government) and planners
- ○Culture and art documentalists (i.e. filmmakers and museum staff)
- ○Emergency respondent planning members and military
- ○Conservation agents (non-natural environment)
- ○Environmentalists of natural environment (public and private)
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

35 [impact_commercialU]

If commercial users had no access to superseded geodata this would create a similar negative impact on society as...

Please choose **only one** of the following:

- ○General public
- ○Archaeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers (government) and planners
- ○Culture and art documentalists (i.e. filmmakers and museum staff)
- ○Emergency respondent planning members and military

- ○Conservation agents (non-natural environment)
- ○Environmentalists of natural environment (public and private)
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

[Question 36 and 37 are equal to question 3 and 4]

Map and remote sensed data scales

38 [comment_scales]

The question was: Imagine two data sets of the same area with the same basic layers: the bigger scale (1:5000) and the smaller scale (1:10000). All versions of the bigger scale are preserved and accessible. How would the following user groups be affected by the systematic elimination of all smaller scale superseded versions?

The following table shows the answers that have been selected by the majority in the first round:

User groups	Effect on the user group if only bigger scale is preserved
General public	would be slightly affected (some affairs could not be resolved any more)
Archaeologists	would be severely affected (could not work or reach their goal anymore)
Historians	would be severely affected / would not be affected
Geographers	would be severely affected (could not work or reach their goal anymore)
Lawyers	would be severely affected (could not work or reach their goal anymore)
Policy makers (government)	would be severely affected (could not work or reach their goal anymore)
Culture and art documentalists (i.e. filmmakers and museum staff)	would be slightly affected (some affairs could not be resolved any more)
Emergency respondent planning members	would be severely affected (could not work or reach their goal anymore)
Conservation agents (non-natural environment. I.e. ancient building inspectors)	would be severely affected (could not work or reach their goal anymore)
Environmentalists of natural environment (public and private)	would be slightly affected (some affairs could not be resolved any more)
Architects and engineers	would be severely affected (could not work or reach their goal anymore)
Institutions that update and maintain geodata	would be severely affected (could not work or reach their goal anymore)
Undergraduate teachers and students	would be slightly affected (some affairs could not be resolved any more)

This result is surprising, taking into account that the eliminated data set (small scale) has less detail than the one that is preserved (bigger scale).

Please comment on why these user groups are affected in this way.

Please write your answer(s) here:

- Social scientists
- Geophysicists
- General public
- Archeologist
- Historians
- Geographers
- Lawyers
- Policy makers (government)
- Culture and art documentalists (i.e. filmmakers and museum staff)
- Emergency respondent planning members
- Conservation agents (non-natural environment. I.e. ancient building inspectors)
- Environmentalists of natural environment (public and private)
- Architects and engineers
- Institutions that update and maintain geodata
- Undergraduate teachers and students

Snapshot intervals or time series

39 [max_time_interval]

The question was: When data sets are regularly updated, like often is the case in GIS, archives can store snapshots. A snapshot is a copy of the state of the data base at a particular moment. At what maximum time interval should the vector map data be archived so that the user groups can still work (reach their goal). In other words: if the time interval between the snapshots in the archive would be longer than that indicated below, the user group would be severely affected.

The intervals are measured in years, even though some data sets are updated more or less frequently.

Use 0 for indicating: "this user group needs all updates".

The following table shows the results from the first round:

User groups	Maximum time interval with which a snapshot should be taken. Median of the estimation of the frist round
General public	5 years

Archaeologists	25 years
Historians	10 years
Geographers	0 = Needs every update
Lawyers	1 year
Policy makers (government)	1 year
Culture and art documentalists (i.e. filmmakers and museum staff)	10 years
Emergency respondent planning members	0 = Needs every update
Conservation agents (non-natural environment. I.e. ancient building inspectors)	10 years
Environmentalists of natural environment (public and private)	1 year
Architects and engineers	1 year
Institutions that update and maintain geodata	0 = Needs every update
Undergraduate teachers and students	5 years

For this second round please diferenciate between time intervals for remote sensed data and vector data bases.

Below, please indicate the maximum time interval of remote sensed data needed by the user groups.

Geophysicists
Social scientists
Commercial users
General public
Archaeologists
Historians
Geographers
Lawyers
Policy makers (government) and planners
Culture and art documentalists (i.e. filmmakers and museum staff)
Emergency respondent planning members and military
Conservation agents (non-natural environment. I.e. ancient building inspectors)
Environmentalists of natural environment (public and private)
Architects, engineers and construction
Institutions that update and maintain geodata
Undergraduate teachers and students

Use 0 for indicating: "this user group needs all updates".

[Question 40 and 41 are equal to question 3 and 4]

42 [time_interval_vector]

Next, please indicate the maximum snapshot interval of vector data bases needed by the user groups.

Geophysicists

Social scientists
General public
Archaeologists
Historians
Geographers
Lawyers
Policy makers (government)
Culture and art documentalists (i.e. filmmakers and museum staff)
Emergency respondent planning members
Conservation agents (non-natural environment. I.e. ancient building inspectors)
Environmentalists of natural environment (public and private)
Architects and engineers
Institutions that update and maintain geodata
Undergraduate teachers and students

[Question 43 and 44 are equal to question 3 and 4]

45 [comment_snapshots]

Please comment on the following statement:

"When it will be possible to reproduce vector databases from the remote sensor data, then there will be no need for producing snapshots of the vector data."

Please write your answer here:

46 [comment_usage_patern]

How do you feel about the following statement:

"Usage patterns should dictate the actual snapshot frequency: If vector data is used infrequently, it may be that no snapshotting is needed at all (i.e. it can be done on demand). If it is frequently used, then this may merit frequent snapshots (even more frequent than for the related remote sensor data)."

Please choose **only one** of the following:

- ○I strongly agree
- ○I somehow agree
- ○Neither agree nor disagree
- ○I somehow disagree
- ○I strongly disagree

Factors that hinder access

47 [lower_access]

The question was: What current trend/s might reduce the access numbers of the user groups to superseded geodata- You could mention trends that affect all or only part of the user groups. (The trends could be technological, legal, organisational, educational etc.)

Below you will find the answers from the first round.

Please rank them acording to their probability of reducing access numbers to superseded geodata. Please rank from the factor that has the biggest negative impact on access numbers (on the top) to the one with the smallest impact (at the bottom).

Please number each box in order of preference from 1 to 8

- INSPIRE-driven necessities to concentrate on cataloguing of active geodata services
- Legal barriers that restrict access (other than liability)
- Missing standardisation
- Missing knowledge or skills of user groups
- Interest shift to more recent data, because there is more and it is more accessible
- Reduction of the use of maps in education
- Difficulty to access highly tecnological geodata
- Concerns about legal liabilities (if users use the wrong version) make institutions to restrict access to previous versions.

48 [overcome_no_access]

Please specify how many years it might take for the following factors, that hinder access, to disappear, to be avoided or to be solved.

Please choose the appropriate response for each item:

	0-5 years	5-10 years	10-20 years	more than 20 years	never
INSPIRE-driven necessities to concentrate on cataloguing of active geodata services	○	○	○	○	○
Legal barriers that restrict access (other than liability)	○	○	○	○	○
Missing standardisation	○	○	○	○	○
Missing knowledge or skills of user groups	○	○	○	○	○
Interest shift to more recent data, because there is more and it is more accessible	○	○	○	○	○
Reduction of the use of maps in education	○	○	○	○	○
Difficulty to access highly tecnological geodata	○	○	○	○	○
Concerns about legal liabilities (if users use the wrong version) make institutions to restrict access to previous versions.	○	○	○	○	○

[Question 49 and 50 are equal to question 3 and 4]

Factors that improve access

51 [higher_access]

The question was: What current trend/s might increase the access rate of these user groups to superseded geodata. You could mention trends that affect all or only part of the user groups. (The trends could be technological, legal, organisational, educational etc.)

Below you find the answers from the first round.

Please rank the following factors that increase superseded geodata access by their inpact on the increase. The factor with the biggest positive impact on the increse at the top and the factor with the smallest impact at the bottom.

Please number each box in order of preference from 1 to 9

- Open (government) data policies and transparency will bring awareness to the availability of superseded data sets
- The use of GIS-related applications increases
- Increased use of open source software
- Central cataloguing encouraged by initiatives like INSPIRE, APEnet/EUROPEANA and national catalogues
- Software becomes more user friendly / esier to use
- Better data visualization makes comparison of supersede data easier
- Increased processing performace and better algorithms and tools for handling large data sets allow better mining and comparison of superseded data.
- Beginners and intermediate users will have greater knowledge about GIS
- Implementation of open data licenses

52 [adress_access_factor]

Please rank the following factors that increase superseded geodata access by how easily they can be addressed or pursued. The factor that is easiest to address or pursue at the top and the most difficult factor to address or pursue at the bottom.

Please number each box in order of preference from 1 to 9

- Open (government) data policies and transparency will bring awareness to the availability of superseded data sets
- The use of GIS-related applications increases
- Increased use of open source software
- Central cataloguing encouraged by initiatives like INSPIRE, APEnet/EUROPEANA and national catalogues
- Software becomes more user friendly / esier to use
- Better data visualization makes comparison of supersede data easier

- Increased processing performace and better algorithms and tools for handling large data sets allow better mining and comparison of superseded data.
- Beginners and intermediate users will have greater knowledge about GIS
- Implementation of open data licenses

Factors that hinder archiving

53 [hinder_archiving]

Please rank the following factors that hinder superseded geodata archiving by their impact.

The factor with the biggest negative impact on archiving at the top and the factor with the smallest impact at the bottom.

Please number each box in order of preference from 1 to 7

- No clear responsibilities inside organizations about the management of geodata systems
- Records management initiatives do not include geodata as main information assets
- Data protection issues / Intimacy laws
- Intellectual property legislation
- Decreasing public funding / budget restrictions
- Concerns about legal liabilities (if users use the wrong version) make institutions to not archive previous versions.
- INSPIRE-driven necessities to concentrate on cataloguing of active geodata services

54 [overcome_no_archive]

Please specify how many years it might take for the following factors, that hinder archiving, to disappear, to be avoided or to be solved.

Please choose the appropriate response for each item:

	0-5 years	5-10 years	10-20 years	more than 20 years	never
No clear responsibilities inside organizations about the management of geodata systems	○	○	○	○	○
Records management initiatives do not include geodata as main information assets	○	○	○	○	○
Data protection issues / Intimacy laws	○	○	○	○	○
Intellectual property legislation	○	○	○	○	○
Decreasing public funding / budget restrictions	○	○	○	○	○
Concerns about legal liabilities (if users use the wrong version) make institutions to not archive previous versions.	○	○	○	○	○
INSPIRE-driven necessities to concentrate on cataloguing of active geodata services	○	○	○	○	○

[Question 55 and 56 are equal to question 3 and 4]

Factors that improve archiving

57 [favor_archiving]

The question was: Mention one or more measures that could improve, compensate or prevent the above mentioned situation of geodata archiving.

Below you will find the answers from the first round.

Please rank the following factors by their capacity to improve archiving of superseded geodata. The factor with the biggest positive impact on archiving at the top and the factor with the smallest impact at the bottom.

Please number each box in order of preference from 1 to 11

- Disminuishing the difference with other data subject to archiving by integrating geodata in organization`s general preservation and records management policies
- Reduce cost by controlled disposal of superfluous data
- Prevention of disaster (risk management)
- Creation of value by re-use of superseded geodata
- Legal efforts on a global scale
- Standardisation efforts on a global scale
- Presentation of important historical facts, stories, research in the media, along with the notion that without archives, this work would not have been possible.
- Include archiving in the production process
- Clear metadata on the data to ensure users can identify and understand the data version.
- Finding cheaper ways of archiving
- Assuring that the archive is well protected to built trust in its capacity to protect intimacy law.

58 [adress_archiving_f]

Please rank the following factors that improve archiving of superseded geodata by how easily they can be addressed or pursued. The factor that is easiest to address or pursue at the top and the most difficult factor to address or pursue at the bottom.

Please number each box in order of preference from 1 to 11

- Disminuishing the difference with other data subject to archiving by integrating geodata in organization`s general preservation and records management policies
- Reduce cost by controlled disposal of superfluous data
- Prevention of disaster (risk management)

- Creation of value by re-use of superseded geodata
- Legal efforts on a global scale
- Standardisation efforts on a global scale
- Presentation of important historical facts, stories, research in the media, along with the notion that without archives, this work would not have been possible.
- Include archiving in the production process
- Clear metadata on the data to ensure users can identify and understand the data version.
- Finding cheaper ways of archiving
- Assuring that the archive is well protected to built trust in its capacity to protect intimacy law.

Functionalities users are looking for

59 [expectations_archive]

The following information, qualities or functionalities might be important to some user groups regarding a geodata archive and superseded data:

- Authenticity of data (information about eventual changes the data has gone through since it is archived)
- Integrity (to be sure the information is not altered)
- Accuracy (to know the accuracy of the data)
- Precision (to know the precision of the data)
- Completeness (to know the degree of completeness of the data)
- Online visualization of the superseded data
- Free access to superseded data (at no cost)
- Information about the authorship of the data
- Information about the creation and change dates of the data
- Information about lineage (knowing how data was created: instruments and calibration,

Can you think of other expectations or qualities that could be important to some user groups? If so, please comment below.

Please write your answer here:

Thank you very much for your input!

- software and parameters)
- Links to related data
- High availability (Short time between request and download possibility)
- Capacity to combine with other data sets
- High colour depth of data
- High resolution
- Homogeneity between series

- Homogeneity within the time series
- Availability of the software to read the data

Annex 3: Online Delphi survey (group two, second round)

2nd round of the Delphi study about developments in GIS and file formats

Welcome to the second round of questions for the Delphi study on digital maps and file formats.

The goal of this second round is to give you feedback on the answers of the other experts and to narrow the deviation in the group answers in order to reach a more precise estimation.

I would also like to emphasize that for the sake of this study I use the term geodata only for data that is used to create maps. I do not include satellite, seismic or atmospheric data.

Please remember the following definitions used in the study:

Vector data base : the vector content of a data base or GIS.

Vector map : the vector data in a processed form, represented as a map and in a sharable vector file format. An ESRI shapefile is a vector map.

Geospatial data = geodata: vector and raster data needed for creating maps. This includes remote sensed data, elevation models, data bases, vector and raster maps but excludes thematic tabular data for the sake of this study.

Version : every time a map layer or data base content is updated a new version is created. In some cases the new version overwrites the old information but modern GIS can associate a date to each update so that a sequence of versions is created and the historic development can be reconstructed. In remote sensing a new version is created when the same area has been photographed (or sensed) at the same scale for a second or subsequent time. For edited maps the new version is the new updated edition.

Superseded data: all versions that are not the most current. I prefer to talk about superseded data rather than historic, because there is a certain expectation for historic data to be antique. This study covers all ages of superseded data: the data that just became superseded one week ago as well as the data that is more than 100 years old.

Snapshot: a snapshot is a copy that represents the state of a data base at a certain moment in time.

Time series: several versions of the same data set form a time series. Every version is created at a different moment in time, usually at regular intervals, therefore the name.

There are 50 questions in this survey

User groups for geodata

1 [present_users]

These are the user groups we used in the first round of this study:

- Historians
- Geographers
- General public
- Archaeologists
- Policy makers (Government)
- Architects and geo-related engineers
- Cultural and arts documentalists (i.e. filmmakers or museum staff)
- Lawyers
- Conservation agents of the non-natural environment (I.e. ancient building inspectors)
- Environmentalists of the natural environment
- Emergency response planners
- Undergraduate teachers and students
- Institutions that maintain and updata geodata (i.e. national survey institutions)

The following additional groups were mentioned in the first round: social scientists, seismologists, geologists, tourists, journalists, (urban) planners, defence/military, ONGs (human development), statisticians, professions affected directly by the weather (fishermen etc.), outdoor athletes, utility companies, retailers, banks, insurances, construction companies.

Journalists, ONG's and statisticians will join the new group of "Social scientists".

Seismologists and geologists will be in the new user group of "Geophysicists".

Tourists and outdoor athletes are included in the "General public".

Planners and urban planners will join the group of "Policy makers".

Defence and military will join the group of "Emergency respondent planning teams".

Users of weather related data will be excluded, as this study concentrates on data that is used for map creation.

Retailers, banks, utility companies and insurance companies will form the new group of "Commercial users".

Regarding App-users; those who create them or publish geodata through them will be regarded as "Commercial users" those who download and use the apps will pertain either to "General public" or to one of the other professions.

Finally, construction companies are to be included in the group of "Architects and geo-related engineers".

This results in the following new user groups:

- Geophysicists (including seismologists, geologists)
- Social Scientists (including economists, journalists, statisticians, people involved in development)
- Commercial users (including banks, utility companies, retailers etc.)

2 [familiar_with]

How familiar are you with the following user groups?

Please choose the appropriate response for each item:

	I have no experience with this user group	I have some experience with this user group	I have a lot of experience with this user group
Geophysicists (including seismologists and geologists)	○	○	○
Social scientists (statisticians, journalists, human development etc.)	○	○	○
Commercial users (banks, utility companies, insurances etc.)	○	○	○
General public (tourists, outdoor sports etc.)	○	○	○
Archeologists	○	○	○
Historians	○	○	○
Geographers	○	○	○
Lawyers	○	○	○
Policy makers (government) and planers	○	○	○
Culture and art documentalists (i.e. filmmakers and museum staff)	○	○	○
Emergency respondent planning members and military	○	○	○
Conservation agents (non-natural environment. I.e. ancient building inspectors)	○	○	○
Environmentalists of natural environment (public and private)	○	○	○
Architects, engineers and construction	○	○	○
Institutions that update and maintain geodata	○	○	○
Undergraduate teachers and students	○	○	○

You could be familiar with those user groups because they are part of your institutions clients, or just because you have some knowledge about their needs.

3 [scales]

Imagine two data sets of the same area with the same basic layers: the bigger scale (1:5000) and the smaller scale (1:10000). All versions of the bigger scale are preserved and accessible. How would the following user groups be affected by the systematic elimination of all smaller scale superseded versions?

Please choose the appropriate response for each item:

	would not be affected	would be slightly affected (some affairs could not be resolved any more)	would be severely affected (could not work or reach their goal anymore)
General public	○	○	○
Archeologists	○	○	○
Historians	○	○	○
Geographers	○	○	○
Lawyers	○	○	○
Policy makers (government)	○	○	○
Culture and art documentalists (i.e. filmmakers and museum staff)	○	○	○
Emergency respondent planning members	○	○	○
Conservation agents (non-natural environment. I.e. ancient building inspectors)	○	○	○
Environmentalists of natural environment (public and private)	○	○	○
Architects and engineers	○	○	○
Institutions that update and maintain geodata	○	○	○
Undergraduate teachers and students	○	○	○

4 [type_of_user]

Please classify the following user groups into these categories:

- frequent users that access or demand big amounts of superseded data
- frequent users that access or demand small amounts of superseded data
- un-frequent users that access or demand big amounts of superseded data
- un-frequent users that access or demand small amounts of superseded data

Please choose the appropriate response for each item:

	frequent user, needs big amounts	frequent user, needs small amounts	infrequent user, needs big amounts	infrequent user, needs small amounts
Geophysicists (including seismologists and geologists)	○	○	○	○
Social scientists (statisticians, journalists, human development etc.)	○	○	○	○
Commercial users (banks, utility	○	○	○	○

	frequent user, needs big amounts	frequent user, needs small amounts	infrequent user, needs big amounts	infrequent user, needs small amounts
companies, insurances etc.)				
General public (tourists, outdoor sports etc.)	○	○	○	○
Archaeologists	○	○	○	○
Historians	○	○	○	○
Geographers	○	○	○	○
Lawyers	○	○	○	○
Policy makers (government) and planners	○	○	○	○
Culture and art documentalists (i.e. filmmakers and museum staff)	○	○	○	○
Emergency respondent planning members and military	○	○	○	○
Conservation agents (non-natural environment. I.e. ancient building inspectors)	○	○	○	○
Environmentalists of natural environment (public and private)	○	○	○	○
Architects, engineers and construction	○	○	○	○
Institutions that update and maintain geodata	○	○	○	○
Undergraduate teachers and students	○	○	○	○

Market share of file formats

5 [current_market]

The question was: Think of the market for digital maps as raster maps, vector maps and databases with vector map information. Estimate the percentage (%) the following file formats have in the current market.

The following table shows the median of the estimation made in the first round.

File format	Estimated percentage in the market of digital maps	Standard deviation
DGN	12,5	6,9
DXF	7,5	7,3
ESRI Shapefile	30	12,1
MrSID	10	12
GeoPDF	3	2,8
MMZ Miramon	0,5	1,3
KML / KMZ	20	13,6
GML	5	10,7
GDB Geodatabase	15	13,4

Please make a new guess based on this knowledge.

Please write your answer(s) here:

- DGN
- DXF
- ESRI Shapefile
- MrSID
- GeoPDF
- MMZ (Miramon)
- KML (KMZ)
- GML
- GDB (ESRI geodatabase)

6 [market_shareOrtho]

The question was: Think of the market of orthocorrected aerial photography. Estimate the percentage (%) the following file formats have in the current market for orthocorrected aerial photography.

The following table shows the median of the estimation made in the first round:

File format	Median of the estimated percentage in the market for orthocorrected images	Standard deviation
MrSID	45	20,3
GeoPDF	10	6,3
GDB Geodatabase	10	19,9
GeoTIFF	40	24,9

Please make a new estimation, based on the knowledge of the group answer.

Please write your answer(s) here:

- MrSID
- GeoPDF
- GDB (ESRI geodatabase)
- GeoTIFF

7 [market_shareDEM]

The question was: Think of the market of digital elevation models. Estimate the percentage (%) the following file formats have in the current market for digital elevation models.

The following table shows the median of the group estimation made in the first round:

File format	Median of the estimated share in the market for digital elevation models	Standard deviation
ASCII ESRI	50	26,8

Please make a new estimation based on the knowledge of the group answer.

Please write your answer(s) here:

- ASCII ESRI (ARC/INFO ASCII GRID)

8 [My_knowledge2]

Regarding the previous questions, how do you feel?

Please choose **only one** of the following:

- ○not confident about your answer
- ○little confidence about your answer
- ○quite confident about your answer
- ○confident about your answer

9 [not_confident]

If you really do not feel confident at all consider passing this question to a collegue or letting the "no answer" option selected.

Only answer this question if the following conditions are met:
° ((My_knowledge2.NAOK == "A1"))

Distribution of vector vs raster maps

10 [vector_raster]

The question was: Think of basemaps. How is the distribution of superseded raster to vector maps that are available online? Please estimate the percentage for both types.

The answers to this question can be divided into two clear groups.

Participants who believe that the bulk of superseded data is in raster format, have given the following arguments:

- Because superseded maps that are digitized are mainly digitized into a raster format.
- Due to companies still being scared of having their data stolen when the data is available in a vector format
- Due to the performance required, often raster formats are favoured.
- Vector formats are more complicated to process and visualize (both for client and server).
- Vector data are not suitable for showing shaded relief / topographic maps.
- Rendering vector maps in a front-end (e.g. browser) is still a niche technology, though it's catching up fast

- Slippy map frameworks such as openlayers or leaflet make it very easy to work with tiled raster basemaps, whereas vector basemaps generally require more bandwidth.
- So many old maps have been scanned. Few vector products are available once superseded.

On the other hand, participants who believe that the bulk is in vector formats, have given the following arguments:

- Popular online web mapping tools are now vector based
- Vector lends itself easier to larger number of applications.
- Use of raster data for analysis is not so generalized.

Please make a new estimation based on these arguments.

Please write your answer(s) here:

- Vector basemaps
- Raster basemaps

Question 11 and 12 are equal to questions 8 and 9.

Increase/decrease use in production

13 [increase_use]

The question was: How might the market share of the following file formats evolve in the production of geodata in the next 10 years.

The following table shows the group estimation made in the first round:

File format	Answer of the majority about the future development in the market
DGN	Stay the same
DXF	Decrease or stay the same
ESRI Shapefile	Decrease or stay the same
ASCI ESRI	Decrease
MrSID	Decrease
GeoTIFF	Increase
GeoPDF	Increase
MMZ Miramon	Not enough answers
KML	Increase
GML	Increase
GDB Geodatabase	Increase

Please make a new guess based on the above information.

Please choose the appropriate response for each item:

	their use in production will increase	stays the same	their use in production will decrease
DGN	○	○	○
DXF	○	○	○
ESRI Shapefile	○	○	○
ASCII ESRI (ARC/INFO ASCII GRID)	○	○	○
MrSID	○	○	○
GeoTIFF	○	○	○
GeoPDF	○	○	○
MMZ (Miramon)	○	○	○
KML (KMZ)	○	○	○
GML	○	○	○
GDB (ESRI geodatabase)	○	○	○

Question 14 and 15 are equal to questions 8 and 9.

Increase/decrease use by consumers

16 [increase_consum]

The question was: How might the market share of the following file formats evolve in the consumption of geodata in the next 10 years.

The following table shows the group estimation made in the first round:

File format	Answer of the majority about the future development in the market
DGN	Not enough answers
DXF	Decrease
ESRI Shapefile	Decrease
ASCI ESRI	Decrease
MrSID	Not enough answers
GeoTIFF	Increase
GeoPDF	Increase
MMZ Miramon	Not enough answers
KML	Increase
GML	Increase
GDB Geodatabase	Increase

Please make a new guess based on the above information.

Please choose the appropriate response for each item:

	increase use by consumers	stays the same	decrease use by consumers

	increase use by consumers	stays the same	decrease use by consumers
DGN	○	○	○
DXF	○	○	○
ESRI Shapefile	○	○	○
ASCII ESRI (ARC/INFO ASCII GRID)	○	○	○
MrSID	○	○	○
GeoTIFF	○	○	○
GeoPDF	○	○	○
MMZ (Miramon)	○	○	○
KML (KMZ)	○	○	○
GML	○	○	○
GDB (ESRI geodatabase)	○	○	○

Question 17 and 18 are equal to questions 8 and 9.

Data sets for science

19 [dataset_science]

The question was: For this question, please think of time series of superseded data sets and their potential interest for science. Please rank the data sets that will interest more scientists.

The following table shows the weighted ranking, taking into account how confident the participants felt with their answers.

Ranking	Data sets
1	Vector data bases (GIS and other data bases)
2	Vector maps (shareable files)
3	Orthocorrected imagery (georeferenced)
4	Digital elevation models
5	Raw lidar data
6	Raw aerial photography
7	Raster maps (shareable files)

Please specify which areas of sciences, 10 years from now, might use superseded time series of these data sets. Please write one or more science topics into the space next to the data sets.

Please write your answer(s) here:

- Raw aerial photography
- Raw lidar data
- Digital elevation models

- Orthocorrected imagery (georeferenced)
- Vector data bases (GIS and other data bases)
- Vector maps (shareable files)
- Raster maps (shareable files)

Please be as specific as possible. The superseded data they use might be older than 10 years.

Data sets for commercial use

20 [commercial_interrest]

The question was: For this question, please think of time series of superseded data sets and their potential interest for commercial use. Rank at the top the data sets that might interest more companies.

The following table shows the ranking made in the first round:

Ranking	Data sets that are of interest for commercial use
1	Vector data bases (GIS and other data bases)
2	Orthocorrected imagery (georeferenced)
3	Vector maps (shareable files)
4	Raster maps (shareable files)
5	Digital elevation models
6	Raw aerial photography
7	Raw lidar data

Which industries or what type of companies might be interested in superseded time series of these data sets in 10 years? Please write into the field next to the data sets.

Please write your answer(s) here:

- Raw aerial photography
- Raw lidar data
- Digital elevation models
- Orthocorrected imagery (georeferenced)
- Vector data bases (GIS and other data bases)
- Vector maps (shareable files)
- Raster maps (shareable files)

Data sets might be older than 10 years.

Technological development

21 [favor_archiving]

The question was: What current technological developments favour archiving versions of superseded geodata?

Below you will find the answers I got from the first round.

How big is the role of these technological developments in archiving versions? Please rank from the developments that play the biggest role in archiving versions at the top to those that have the smallest role.

Please number each box in order of preference from 1 to 14

- Cheaper IT
- Use of cloud computing
- Use of standards by the producer
- Use of standards by the archive
- Compliance with legal incentive to store versions like INSPIRE
- Hybrid open-source vector/raster databases.
- Use of faster interfaces
- Increasing knowledge of basic technology among archivists and (geo-)archive users
- Use of GeoGit
- Use of PostGIS
- Use of NoSQL databases
- Use of linked data
- Increase in storage and retrieval capabilities
- Use of open source software by producers

Rank the technological development that you think would have the biggest positive impact on archiving versions at the top.

Question 22 and 23 are equal to questions 8 and 9.

24 [query_online]

The question was: How many years will it take until at least 50% of the superseded map content that is available online will be queryable for features without having to download it?

The median of all the answers was: 10 years.

With how much confidence you would say that in 10 years at least 50% of the superseded map content that is available online will be able to be queried for features without having to download it?

Please specify the confidence as a percentage.

Each answer must be at most 100

Please write your answer here:

Vector archiving

25 [hinder_archiving]

The question was: What current technological constraints hinder or complicate archiving versions of superseded vector data bases or shareable vector maps?

Below, you will find the constraints you mentioned in the first round. I would like you to estimate how many years it might take for them to disappear or to be overcome.

Please choose the appropriate response for each item:

	0-5 years	5-10 years	10-20 years	more than 20 years	never
Use of proprietary data formats by producers	○	○	○	○	○
Use of non-standard formats	○	○	○	○	○
Standards for (properly) documenting the data models are too complex	○	○	○	○	○
Fast growth of vector data bases and shareable vector maps	○	○	○	○	○
Missing or insufficient metadata or semantics	○	○	○	○	○
The steep learning curve in organizing the various dat files in these data bases and maps.	○	○	○	○	○
Bad compression rate of vector data	○	○	○	○	○
Version control systems (Git etc) are not set up for the large number that would be needed to represent a significant data set due to the underlying file system inode capacity	○	○	○	○	○
Current practice of producers does not include adding temporal metadata	○	○	○	○	○
The focus on GIS as "layers"	○	○	○	○	○
Relational database models	○	○	○	○	○
Difficulty in scaling graph databases	○	○	○	○	○
Lack of appropriate standards for representing vector data with temporal aspects.	○	○	○	○	○

Question 26 and 27 are equal to questions 8 and 9.

28 [measure_vectordb]

The question was: Mention one or more measures that could improve the above mentioned situation of vector database and vector map archiving.

The elements below are the answers from the first round.

Which are the measures with the biggest positive impact on superseded vector data archiving? Please rank from the biggest positive impact (at the top) to the smallest (at the bottom).

Please number each box in order of preference from 1 to 12

□ Simplify the documentation process

□ Use of open source software by producers

□ Use of standard file format by producers

□ Use of standard data models by producers

□ Use of GeoPackage

□ Development of version control software that can handle larger numbers

□ Development of better compression for vector data

□ Teaching producers to organize data

□ Teaching users to understand data

□ Development of file system that can handle massive amounts of small GeoJSON files

□ Increasing use of unstructured databases

□ Development of simple spatial temporal design models

Raster archiving

29 [hinder_imagery]

The question was: What current technological constraints hinder or complicate archiving versions of superseded raster geodata?

Below, you will find the constraints you mentioned in the first round. I would like you to estimate how many years it might take for them to disappear or to be overcome.

Please choose the appropriate response for each item:

	0-5 years	5-10 years	10-20 years	more than 20 years	never
Proprietary and undocumented IMPLEMENTATIONS of the compressed raster standards	○	○	○	○	○
Use of proprietary software by producers	○	○	○	○	○
Use of non-standard formats	○	○	○	○	○
Georeferencing and fitting problems	○	○	○	○	○
Not enough bandwidth	○	○	○	○	○
High cost of storage	○	○	○	○	○
Slow storage technologies	○	○	○	○	○
Bad compression rate of vector data	○	○	○	○	○

	0-5 years	5-10 years	10-20 years	more than 20 years	never
Lack of appropriate workable metadata standards	○	○	○	○	○
Producers don't add temporal metadata	○	○	○	○	○
High volume of data	○	○	○	○	○
Inappropriate analysis tools and methods in current GIS.	○	○	○	○	○

Question 30 and 31 are equal to questions 8 and 9.

32 [measure_raster]

The question was: Mention one or more measures that could improve the above mentioned situation of remote sensed imagery archiving.

The elements below are the answers from the first round. Which are the measures with the biggest positive impact on superseded vector data archiving? Please rank from the biggest positive impact (at the top) to the smallest (at the bottom).

Please number each box in order of preference from 1 to 9

□ Use of open source software by producers

□ Use of standard file format by producers

□ Better compression technology

□ Cheaper storage

□ Faster storage

□ Higher capacity bandwidth

□ Big Data applications

□ Simple workable metadata standards

□ Use of temporal metadata by producers

Functionalities

33 [prevent_lossFunction]

How could GeoJSON help archiving geodata versions and how can it help prevent loss of functionalities?

Please write your answer here:

34 [Drop_down]

To what other user groups are Geophysicists most similar in terms of the functionality they need from the data?

If you think there is no similarity regarding what the user group needs from the data, choose the same user group again from the drop down list.

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

35 [similar_socialSCI]

To what other user groups are Social scientists most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata

- ○Undergraduate teachers and students

36 [similar_commercial]

To what other user groups are Commercial users most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

37 [similar_general]

To what other user groups is the General public most similar in terms of the functionality it needs from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

38 [similar_archeologist]

To what other user groups are Archaeologists most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

39 [similar_historians]

To what other user groups are Historians most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

40 [similar_geographers]

To what other user groups are Geographers most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

41 [similar_lawyers]

To what other user groups are Lawyers most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

42 [similar_policymaker]

To what other user groups are Policy makers most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

43 [similar_cultureArts]

To what other user groups are Culture and Arts documentalists most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

44 [similar_emergency]

To what other user groups are Emergency response teams most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

45 [similar_conservAgent]

To what other user groups are Conservation agents of the built environment most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

46 [similar_environment]

To what other user groups are Environmentalists most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

47 [similar_engineers]

To what other user groups are Architects and geo-related Engineers most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

48 [similar_updateInstit]

To what other user groups are Institutions that update and maintain geodata most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

49 [similar_teachersStud]

To what other user groups are Undergraduate teachers and students most similar in terms of the functionality they need from the data?

Please choose **only one** of the following:

- ○Geophysicists
- ○Social scientists
- ○Commercial users
- ○General public
- ○Archeologists
- ○Historians
- ○Geographers
- ○Lawyers
- ○Policy makers
- ○Culture and art documentalists
- ○Emergency response and military
- ○Conservation agents
- ○Environmentalists
- ○Architects, engineers and construction
- ○Institutions that update and maintain geodata
- ○Undergraduate teachers and students

Snapshots for the archive

50 [no_snapshot_needed]

Please comment on the following statement:

"10 years from now, it will be possible to reproduce vector databases from the remote sensor data, then there will be no need for producing snapshots of the vector data for the archive (it can be done on demand)."

Please write your answer here:

Thank you very much for your input!

Annex 4: Semi-structured interview guide for the focus groups

Presentarme.

Explicar el objetivo de la reunión.

Distribuir y explicar la carta de consentimiento.

Encender la camera.

Explicar las tres partes de la entrevista:

He dividido la reunión en tres partes. En la primera me gustaría que cada uno se presente rápidamente y que me explique un proyecto o un contexto en que está utilizando documentación histórica, o mapas históricos o fotografía aérea histórica. Luego en la segunda parte me gustaría que me expliquéis que sería el archivo digital ideal: qué contendría, hasta qué año atrás iría. Qué ficheros habría, qué formatos? En la tercera parte, os explico la selección que podría haber en el archivo. Voy exagerar un poco, y me decís: con esta selección podemos seguir trabajando en los proyectos que tienen en mente o se han explicado al inicio, o no. Si es demasiado radical la selección y no se podría trabajar.

1ª parte

Cada uno se presente y explique su uso de información geográfica histórica.

Preguntar por la frecuencia de uso y la cantidad de datos necesitados.

2ª parte

El archivo ideal. Imaginar la carta a los reyes. En conjunto hablamos del archivo ideal.

Aspectos a tratar: Tipos de producto, formatos, años, cobertura del terreno.

3ª parte

La selección que se podría aplicar en un archivo. Podríais seguir trabajar si…

La escala más de detalle se guarda y el resto se borra. ¿Podríais seguir trabajando?

Se guardan los datos cada 5 años, por ejemplo el año 35, 40 y 45. Los años intermedios se borran.

Dentro de una cadena de producción de una cartografía hay muchos productos: la foto, la orthocorregida, los datos, el mapa final, además existen modelos de terreno, cual es de lo que más prescindir? (¡Adaptar pregunta al conocimiento de los participantes!)

Si no se ha agotado volver a mencionar los formatos y las edades de los datos que interesan los usuarios.

Eventualmente: preguntar por los perfiles de los participantes.

Annex 5: Letter of consent for participants of the focus groups

Carta de consentimiento para la participación en un grupo focal

Objetivo:
Usted aceptó participar en un grupo focal (focus group) en el marco del programa de doctorado "Información y documentación en la sociedad del conocimiento" de la doctoranda **Anita Locher**. Su participación ayudará a tomar decisiones en el proceso de diseño de un servicio de archivo histórico de datos geográficos. Esta actividad está pensada para ayudar al diseño del archivo y de ninguna manera pretende evaluar su conocimiento, sus actitudes o comportamientos.

Proceso/transcurso de la actividad:
Usted participará en un grupo focal (entrevista en grupo) junto a 2-5 otras personas con perfil similar que durará entre **1 hora y 1 hora y media**. Se le preguntará sobre su percepción de lo que sería un sistema "ideal" de archivos históricos de datos (funcionalidades, información que tiene que contener, ayudas que tiene que ofrecer, etc). Se le presentará servicios de geodatos recientes e históricos y se le preguntará si le sirven este tipo de datos para el trabajo u ocio. Finalmente se le preguntará sobre pegas y restricciones del sistema que más le impedirían trabajar. Mientras converse con sus colegas estará siendo **filmado y grabado**.

Confidencialidad:
La grabación será visualizada únicamente por la doctoranda. No se hará difusión de la grabación por internet ni cualquier otro medio. La única excepción puede ser una eventual visualización por el tribunal evaluador de la doctoranda a fin de verificar la calidad de la tesis. Para garantizar la confidencialidad, no se relacionará su nombre o datos personales con sus opiniones y grabaciones.

Derecho de desistimiento
Usted tiene derecho de hacer una pausa o de retirarse del estudio en cualquier momento.

Si acepta los términos explicados arriba, por favor, firme abajo:

Nombre y apellido:____

Firma:____

Lugar y fecha:____

Firma de la doctoranda, Anita Locher:_______________________________________

Para más información o cualquier tema relacionado con el proyecto se puede dirigir a: anitalocher@gmail.com

Annex 6: Relation of users who participated in the questionnaire and focus group interviews

Users who responded the initial questionnaire

User identification	Background	Type of his or her research	Gender	Participated in
S1	No answer	Geography	male	Questionnaire
S2	History	Social history / Archaeology	male	Questionnaire
S3	No answer	Archaeology	female	Questionnaire
S4	Geography	Geography	male	Questionnaire
S5	Geography	Geography	female	Questionnaire
S6	Art	Economic and social history	male	Questionnaire
S7	Architecture	Geography and art	female	Questionnaire
S8	Geography	Geography	male	Questionnaire
S9	History	Geography	male	Questionnaire
S10	Geography	Geography and economic and social history	female	Questionnaire
S11	Geography	Geography	male	Questionnaire
S12	Geography	Geography and social history	female	Questionnaire
S13	Geography	Geography	female	Questionnaire

Users who participated in the focus group interviews

User identification	User type	Gender	Participated in
G1_P1	Geographer	female	Focus group
G1_P2	Geographer	male	Focus group
G1_P3	Geographer	male	Focus group
G1_P4	Historian	male	Focus group
G2_P1	Geographer	male	Focus group
G2_P2	Urban planner	male	Focus group
G3_P1	Urban planner	female	Focus group
G3_P2	Urban planner	male	Focus group
G3_P3	Urban planner	female	Focus group
G4_P1	Urban planner	female	Face to face interview
G5_P1	Historian	male	Focus group
G5_P2	Historian	female	Focus group
G5_P3	Historian	male	Focus group
G6_P1	Geographer	male	Focus group
G6_P2	Geographer	male	Focus group
G6_P3	Geographer	male	Focus group

G6_P4	Geographer	female	Focus group
G7_P1	General public	male	Email interview
G8_P1	Urban planner	male	Focus group
G8_P2	Urban planner	female	Focus group
G8_P3	Urban planner	male	Focus group
G9_P1	General public	male	Focus group
G9_P2	General public	male	Focus group
G9_P3	General public	female	Focus group
G10_P1	General public	male	Focus group
G10_P2	General public	male	Focus group
G10_P3	General public	male	Focus group
G10_P4	General public	female	Focus group
G11_P1	Geographer	male	Focus group
G11_P2	Geographer	female	Focus group
G11_P3	Geographer	male	Focus group
G12_P1	Historian	male	Email interview
G13_P1	Geographer	male	Email interview
G13_P2	Geographer	male	Email interview
G14_P1	General public	male	Email interview