Network Engineering Department
Universitat Politècnica de Catalunya

Ph.D. Thesis

Contributions to efficient and secure exchange of networked clinical data

- The MOSAIC System -

Author: Magí Lluch-Ariet

Ph.D. Advisor: Josep Pegueroles-Vallés
Ph.D. co-Advisor: Francesc Vallverdú-Bayes

June 2016

"Share your knowledge. It is a way to achieve immortality"
Dalai Lama XIV

Contact details of the author
Magí Lluch-Ariet
e-mail: magi.lluch@gmail.com
Researchgate: www.researchgate.net/profile/Magi_Lluch-Ariet

Front page design by Roger, Griselda and Guifré Lluch-Sabartés

To the sharing lovers

Abstract

The understanding of certain data often requires the collection of similar data from different places to be analysed and interpreted. Multi-Agent Systems, interoperability standards (DICOM, HL7 or EN13606), clinical Ontologies and coding standards (SNOMED-CT, ICD10 or ATC) are facilitating data exchange among different clinical centres around the world. However, as more and more data becomes available, and more heterogeneous this data gets, the task of accessing and exploiting the large number of distributed repositories to extract useful knowledge becomes increasingly complex. Beyond the existing networks and advances for data transfer, specific data sharing protocols to support multilateral agreements can be useful to exploit the knowledge of distributed Data Warehouses. The access to a certain data set in a federated Data Warehouse may be constrained by the requirement to deliver another specific data set, and when bilateral agreements between two nodes of a network are not enough to solve the constraints for accessing to a certain data set, multilateral agreements for data exchange can be a solution.

The research carried out in this PhD Thesis comprises the design and implementation of a Multi-Agent System for multilateral exchange agreements of clinical data, and evaluate how those multilateral agreements increase the percentage of data collected by a single node from the total amount of data available in the network. Different strategies to reduce the number of messages needed to achieve an agreement are also considered.

The results show that with this collaborative sharing scenario the percentage of data collected dramatically improve from bilateral agreements to multilateral ones, up to reach, in the specific evaluation scenario used, almost all data available in the network.

Keywords and Categories

Keywords that define this Thesis are: Data Sharing, Clinical Data Sharing, Data Exchange, Network Protocol, Multi-Agent System, Data Warehouse, Federated Data Warehouse.

The categories related to this Thesis research based on the ACM Computing Classification Scheme are: Network Protocols (C.2.2), Distributed Systems - Distributed databases (C.2.4), Distributed Artificial Intelligence - Multiagent systems (I.2.11), Systems and Software - Distributed Systems (H.3.4) and Online Information Services - Data sharing (H.3.5).

Acknowledgements

I would like to specially thank Prof. Josep Pegueroles-Vallés that coordinated the scientific advances of this research and provided the top level guidelines to boost its results, together with Prof. Francesc Vallverdú-Bayes. Thanks also to my colleagues from the Network Engineering department at UPC. My special gratitude to Albert Brugués for his contribution in programming tasks and David Rebollo for his review of some mathematical aspects. Thanks also to my colleagues from the Personalised Computational Medicine research line and the whole R&D eHealth Department at the Eurecat Technology Centre, and also to my colleagues in the Synergy-COPD and HealthAgents EU projects for the inspiring discussions I had with them. Thanks also to all people I had the chance to meet and share the ideas of this Thesis that directly or indirectly inspired me to find new ways to contribute to the data sharing scenario and worldwide collaboration. Thanks also to Vicent Ribas and Ignasi Belda for their technical review and to my friends and family for their strong support and understanding of the efforts devoted to this research during holidays and weekends.

This research has been partially funded by the project TAMESIS (TEC2011-22746) and the Synergy-COPD research grant, under the Seventh Framework Program of the European Commission as a Collaborative Project with contract no.: 270086 (2011-2014).

Magí Lluch-Ariet
December 2015

Contents

Part I

Introduction

Chapter 1

Reasons for Efficient and Secure Exchange of Networked Clinical Data

"We are at the start of a collaborative revolution
that will be as significant as the Industrial Revolution"
Rachel Botsman

1.1 Collaboration, sharing and exchange

Worldwide collaboration is a fact in most areas of activity. Beyond traditional economic transactions and multi-party agreements, innovative models that join forces to boost collaboration are being deployed. In some cases, this collaboration is based on the concept of sharing (e.g. car sharing) and in others on the exchange of a good or service (e.g. temporary home exchange for accommodation and time banking). Collaboration consumption (sharing goods instead of buying) was selected by the TIME magazine in 2011 as one of the 10 ideas that will change the world [15].

This collaboration is facilitated when the items to share or exchange correspond to knowledge or information, easily transmitted digitally. Public regulations push to open up and share data, for scientific publications [24] (Open Access) and for public data [25, 23] (Open Data). The European Union Open Data Portal [38] and the OpenAIRE project [85] are implementation examples of these policies. Local repositories of data are growing and growing and the emergence of the Big Data era is putting us in front of unmet challenges to exploit the vast amount of data generated.

The amount of data that can be accessed in a network through bilateral exchange agreements can be extended increasing the number of collaborating parties and creating multilateral agreements. Although networks, standards for data representation and communication protocols, are connecting the growing number of worldwide distributed repositories, the search of valuable data and the negotiation of multilateral agreements for data exchange becomes very

complex or impossible to be managed manually. MOSAIC, the protocol presented here, aims to contribute to this objective by facilitating the building process of multilateral agreements for data exchange.

1.2 Healthcare as a scenario for collaboration

Similarly to other fields, the future of medicine is facing ambitious challenges that can not be achieved without the combined analysis of the growing amounts of data sets generated. In spite of the barriers for health data sharing [104] and the regulations for privacy protection, there are strategies to overcome them [79] and MOSAIC aims to be a new instrument in order to facilitate this collaboration.

There is a number of reasons for efficient and secure exchange of networked clinical data that can be illustrated with the following two examples that facilitate both professional and social collaboration.

Professional collaboration
Clinicians often need to compare the information collected from the experiments performed to their patients with information from similar patients in other places. This is needed for accurate diagnosis, prognosis, theragnosis, effective management of the diseases and efficient use of drugs.

Ethical and legal regulations that apply to personal data [22, 26], and the associated data access authorisations to be provided by the ethical committees, must be integrated in any negotiation process for data exchange. Under the assumption that all of this is fulfilled, a clinician may also add some additional constraint and give access to the data only if another dataset is given. However, bilateral agreements between two clinical centres will not always solve those constraints and involving a set of centres in multilateral agreements for data exchange would increase the amount of data potentially accessible in a network.

Social collaboration
We can also imagine a near future with a considerable number of individuals with their own Electronic Health Record (EHR) in their healthcare ID cards including their genotype. That information will be preserved, protected and regulated by law as personal and private information. Nowadays, most data of genetic sequences is collected managed and stored in research labs for scientific purposes, but as the next generation sequencing technology becomes more affordable, beside those big data repositories, new distributed databases are appearing. In a near future, a significant number of this information will be hosted and managed by every individual, and some of them will be reluctant to allow its storage and use in a database. In such scenario people may what to share its own data only if this brings some benefit to them. Someone may be interested to explore the network to find individuals with similar genetic profile, suffering similar illness to learn from their experience and treatment evolution. Similarly to

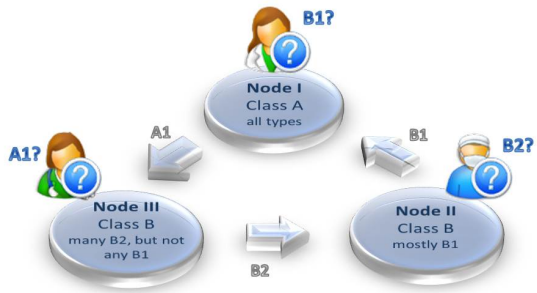

Figure 1.1: Example of a multilateral agreement and data exchange among three nodes

the previous scenario, a person may give access to his personal information only if some other information is given by someone else, and bilateral agreements between two persons may not always solve those constrains. Involving a group of people in multilateral agreements for data exchange would also increase the amount of data potentially accessible in the network.

1.3 Illustrating a simple case of multilateral data exchange

The following example illustrates how a system for clinical data exchange may work. Consider a clinical framework in which patients are classified in different classes (A, B, C, ...) and for each class there is a set of possible diagnoses (A1, A2, ...).

Node-I Hosts a Data Mart (local database of the node, part of the federated data warehouse of the whole network) of cases 'class A' (including all possible diagnoses in the class). Exceptionally a new patient 'class B' needs to be diagnosed and the clinician wants to compare it with some other cases (class B) from some external nodes, already diagnosed as B1.

Node-II Hosts a Data Mart of patients 'class B' with an exceptional considerable number of cases diagnosed as B1. A new patient 'class B' needs to be diagnosed and the clinician wants to compare it with some other cases from external nodes already diagnosed as B2.

Node-III Hosts a Data Mart composed by data of class B, with an exceptional considerable number of cases diagnosed as B2, but without a single case diagnosed as B1. A new case 'class A' needs to be diagnosed and the clinician wants to compare it with other cases already diagnosed as A1 in other centres.

There is not any bilateral agreement for data exchange from the three nodes that solves the data access needs. However, as shown in Figure 1.1, the agreement is possible as follows: i)

Node-III gives the access rights to Node-II for accessing to the cases already diagnosed as B2; ii) Node-II gives access rights to its cases diagnosed as B1 to Node-I and i) Node-I gives access rights to its cases diagnosed as A1 to Node-III.

The process to achieve multilateral agreements involving a large number of nodes in big networks can be facilitated by an automatic process that solves the complex search of the candidate nodes for the possible agreement, solving the needs of data access and providing the access rights to the right nodes.

This thesis presents MOSAIC, a Multi-Agent System and its associated network protocol, that fulfills these requirements and facilitates the achievement of multilateral agreements involving the data exchange among a number of nodes of a network in order to get access to the desired datasets.

Chapter 2

Objectives and Methodology

"Plans are only good intentions unless they immediately degenerate into hard work"
Peter Drucker (1909 - 2005)

2.1 Challenges and main objectives

The main objective of MOSAIC is to get access rights to the data hosted in the Data Marts of a federated data warehouse through multilateral agreements for data exchange. To this end, the MOSAIC protocol will support the negotiation process among the nodes of this network.

The framework in which MOSAIC will operate consist of: i) A network of thousands of nodes, worldwide distributed; ii) A relatively large amount of data to transfer; iii) A complex network of dependencies among the nodes; iv) A strict and heterogeneous legal and ethical policy framework; and v) A preliminary subjective quantification of the value of the data.

The strategies for clinical data exchange among the nodes of a network designed in MOSAIC will provide the following main advantages, that correspond to the main Thesis objectives:

1. Select the right nodes among all candidates to build the best paths possible for the multilateral agreements
2. Minimize the number of messages of the protocol used to achieve a multilateral agreement
3. Protect the data and grant its delivery to the right recipient
4. Increase the amount of data that can be collected with bilateral data exchanges through new multilateral agreements

The main problem of MOSAIC in order to find multilateral agreements includes the problem of finding the shortest path in a complex network. This could be rapidly solved using the Dijkstra algorithm [30]. However, this is possible only if the links between the nodes are known as well as the topology of the whole network. In many scenarios the information of the network

topology is neither available nor complete. As an example of this in the healthcare scenario, a clinician may accept to publish the reference of which datasets are available from his local repository, but the specific permission to allow access to them may not be provided before an explicit data access request from a specific centre is received. Thus, a centralised approach to solve this problem is not feasible and a distributed and dynamic mechanism for the exploration of the paths associated to possible multilateral agreements is needed.

Dissemination and exploitation

An objective of the Thesis is also to publish and present the research results in scientific journals and conferences. The presentation of the initial idea to the scientific community in some international Workshop is the first step, allowing to share views and receive suggestions from other colleagues working with this topic. A publication in a JCR journal is also an objective of this research. For this, Open Access publications are the first options considered as a mean to facilitate the awareness of the Thesis results.

The deployment and use of MOSAIC in the real world is an objective that goes beyond the scope of this Thesis. However, the inclusion of these objectives in new proposals for research and innovation grants are also in the plans of the author, as well as the possibility to fill a patent application to exploit the MOSAIC opportunities in certain fields whose use can provide significant advantages.

2.2 Methodology and development process

Based on the analysis of current systems for data sharing and exchange in the clinical framework, this Thesis provides a design of the MOSAIC System that goes beyond current state-of-the-art. A realistic dataset is prepared for simulating the behaviour of MOSAIC and an implementation to simulate the Agents in a real scenario is built. A theoretical model to predict the impact and performance of MOSAIC is designed and validated through the execution of MOSAIC in a simulated environment.

The design and implementation of MOSAIC faces a number of issues that have to be addressed, namely: i) Design the architecture of the system to support the protocol; ii) Design the protocol architecture, its components and the negotiation strategy; iii) Design the management system of the protocol and iv) Implement and validate the MOSAIC system and protocol.

The Architecture of the MOSAIC System

The first step towards the development of MOSAIC is the design of its basic infrastructure, which includes the selection of the Agent Platform, the DBMS and the Web and SSL servers.

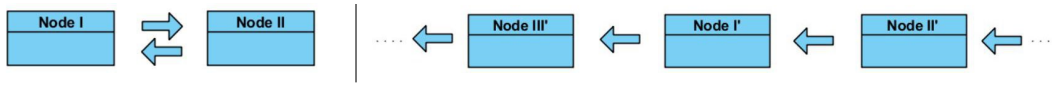

Figure 2.1: Bilateral vs Multilateral agreements. On the left, Node I and Node II built a bilateral agreement for data exchange (Node I delivers its data to Node II and collects the requested data also from Node II). On the right, Node I' collects data from Node II', but delivers its data to Node III', showing a link of a multilateral agreement

A FIPA compliant protocol will be selected for the communication among the Agents. The system will provide the framework for a proper communication among the agents, independent of the Agent platform and a specific messaging library for the agent's use, including the functionalities for the data publication and retrieval. The security of the data hosted in the network and transferred between its nodes must be also granted. This includes the consistency, confidentiality and authenticity of the data.

The System Management

The second step has to face the issues of the global management of the system. This will be based on a directory service that will store the information about the nodes of the network, the active agents, the data they offer, and provide a confident framework for the user identification, validation and data access authorisation.

Part of the global management is also the communication and interface between the system and the user, when this is required. This includes two main aspects: One related to the final authorisation of any agreement that could be made automatically, but the user may also prefer to validate the proposed agreement from the system, in which case the system management will provide the interface for the communication between the agents and the users, facilitating also the confirmation or rejection of constraints before executing any agreement. The other global management aspect, is about the legal framework under which MOSAIC will operate (different from country to country) and the ethical principles specific to each participating centre. The global management system should monitor the fulfillment of both the legal and ethical frameworks, checking and preventing any violation by any possible agreement.

The MOSAIC Communication Protocol

The third step towards the MOSAIC development is the design of the Agents of the protocol, their dialogs and negotiation strategy. In summary, the MOSAIC Agents include i) Agents to publish data available to be exchanged, ii) Agents to request and collect data from the network, and iii) Agents to manage the access to each Data Mart.

The MOSAIC protocol has to identify which nodes have the data requested. After that,

the protocol needs to look for the nodes that can access to that data, directly with bilateral data exchange, or indirectly with multilateral data exchange (see Fig. 2.1). After having the candidates for the agreements, the cost to access to the data can be calculated according to the metrics involved in the quantification of the data value, and a subset of the agreements identified. When more than one agreement is possible, a set of rules to select the best option can be applied.

These are the main actions that the MOSAIC protocol has to support (with references to the *simple case of multilateral data exchange* included as example in section 1.3, Figure 1.1):

- **Publication of the reference of the data available in a node**
  This will make possible to know which data is available in the network (e.g. Node-I notifies the availability of cases from its Data Mart, class A)

- **Transmission of the request for data access**
  A node requesting data must inform which data is looking for, and activate the search (e.g. The clinician in Node-II activates the search of cases diagnosed as B2 in the network)

- **Transmission of the authorisation and access rights to the data**
  A node providing data will have to authorise or reject the access to its data (e.g. Node-III gives access rights to Node-II for accessing to its cases classified as B2)

- **Delivery of the data**
  The protocol will provide data transfer capabilities according to the access right policies. (e.g. Node-III will send the cases classified as B2 to Node-II)

- **Acknowledge the reception of the data**
  The reception of the data has to be confirmed (e.g. Node-II sends the acknowledge of receipt to Node-III when the cases diagnosed as B2 are received)

- **Withdraw the access rights** (if the agreement is not completely fulfilled)
  Partial agreements during the negotiations are not definitive before achieving a complete agreement (e.g. Considering the request launched by Node-II for accessing to cases available in Node-III, access rights given by Node-II to Node-I have no sense if the access rights request for accessing to Node-I from Node-III are rejected)

- **Start, Commit or Rollback the whole transaction**
  Partial transfers of data may not be acceptable if the whole data transfer corresponding to an agreement is not achieved (e.g. In spite of a complete agreement, if when executing the transfer of data, Node-III does not transfer the cases diagnosed as B2 to Node-II, all the data transferred among the other nodes involved in the multilateral agreement should no be used and deleted at the recipient nodes).

The MOSAIC validation

The validation and evaluation of the MOSAIC System will prove first, its correct implementation. This will be tested using a network where all possible bilateral agreements will be created (without MOSAIC) and based on them the total number of cases collected will be measured. Then, MOSAIC will be executed in the same network setting the parameters to limit the length of the exploration paths to two nodes, and the final figures of cases collected using both methods will be compared. The second validation will check the utility of MOSAIC demonstrating the potential increase of data collected by a node in a network comparing the results obtained when the node negotiates the data exchange through bilateral agreements and when the node uses multilateral agreements to get access to the desired data. Finally, two exploration methods for the path exploration will be also tested, comparing their results and demonstrating the flexibility of MOSAIC to adapt its behaviour to different strategies.

For a complete analysis and assessment of MOSAIC, a set of metrics will be identified and tested in a realistic scenario, measuring the improvement of the scores of those metrics. A mathematical equation to calculate the theoretical number of agreements that can be achieved will be also formulated and validated in the evaluation scenario.

2.3 Brief summary of the content

This document is structured in five parts containing nine chapters in total. Part I aims to provide a summary of the reasons and motivational aspects that inspired the idea of the Thesis (Chapter 1), together with and overview of the objectives, main challenges and methodology followed during the Thesis development (Chapter 2). Part II (Chapter 3) is to show the result of the State-of-the-Art analysis in clinical data sharing, highlighting the new contribution provided by MOSAIC. Part III (Chapters 4-7) compiles the key results of the Thesis. In Chapter 4 the architecture and main components and Agents of MOSAIC are described. In Chapter 5, the communication protocol among the MOSAIC Agents and the negotiation process among them are described in detail. In Chapter 6 the Key Performance Indicators that will allow to assess the MOSAIC System are identified, and a predictive model that illustrates the expected improvements in terms of data collected is formalised. Chapter 7 details the results of MOSAIC's execution in a simulated scenario and validates the hypothesis formulated in Chapter 6. The last two parts of the document (Part IV and V) summarise the main achievements and open challenges (Chapter 8) and list the publications and presentations in scientific events resulting from the Thesis research (Chapter 9).

Part II

State of the Art

Chapter 3

Clinical and Biomedical Data Sharing

"Research is to see what everybody else has seen, and to think what nobody else has thought"
Albert Szent-Györgi (1893-1986)

3.1 Clinical and biomedical data

The healthcare sector is especially rich in data, and the growth comes both from digitising existing files and from generating new forms of data. The volume of worldwide healthcare data is estimated to grow from some hundreds of petabytes during the recent years up to tens of exabytes by 2020 [54, 51, 47, 48]. In particular, genomic data is expected to increase at a high rate in the next few years, due to reduced DNA sequencing cost (see Fig. 3.1), that is approaching to 1.000 EUR per whole genome sequencing. The grow of other forms of biomedical data, namely Magnetic Resonance Imaging (MRI), electrocardiograms (ECG), spirometry readings, among others, is feeding EHR and allowing combined analysis of the patient's health status, e.g. 8.000 MRI data sets are already shared and available online [91]. Cells, tissues and biological samples are also key for biomedical research. Their collection and

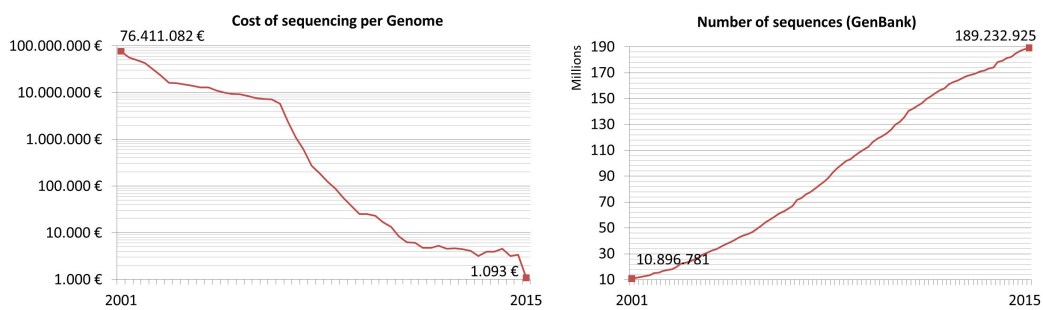

Figure 3.1: Cost and number of genome sequencing. Figures provided by the National Human Genome Research Institute at NIH (USA) [84, 8]

storage in biobanks are emerging worldwide and infrastructures to harmonise how those samples are stored, how they can be accessed and how related they are, are required.

The European Biobanking and Biomedical Resources Infrastructure (BBMRI-ERIC [37, 64, 110]) is one of the largest research infrastructure initiatives in Europe, that provides access to biobanks for the European research community. It's a growing network funded on December 2013, supported by 16 EU Member States and composed by more than 500 biobanks. On November 2015 BBMRI-ERIC is involved in 10 projects (IMI, FP7 and H2020) with ADOPT BBMRI-ERIC as the most relevant to accelerate the adoption of this infrastructure at EU level.

With the use of mobile Health Apps, patients and citizens are not only building social networks, but also collecting and sharing data of three categories: health and fitness tracking data, ii) patient monitoring data (for the management of chronic conditions) and iii) medical examination data (ECG, blood and urine tests, ...). In 2014, more than 100.000 Apps were available and 71% of them with the possibility to import or export health data [93]. In 2015 this figure grew up to more than 165.000 Apps.

Healthcare is characterised by systems that have foundations in national structures such as national or regional health care organisations, and the data generated is usually not open or shared for privacy and data protection reasons. Simultaneously, research data in the area of genetics, genotype to phenotype connection and studies of genetic diseases is generally characterised by international, open, accessible and reasonably comprehensive research databases. Policies on research data sharing are just being published (e.g. The Genomic Data Sharing (GDS) Policy [83] from NIH) providing a legal framework for worldwide collaboration.

The availability of MRI, NGS, biological samples from biobanks, along with the other forms of biomedical data linked with the health status of every individual will radically change the biological, clinical and pharmaceutical research, and the future of medicine. A proper stratification of patients combined with the knowledge from others previous cases will allow to give more precise diagnoses and personalised prognosis and therapeutic plans. To this end, the vast amount of distributed diverse and evolvable data marts have to be connected.

The advances on Personalised Medicine for diagnosis, prognosis and theragnosis, require the combined analysis of similar cases, connecting people in huge networks. The exchange and sharing of clinical and biomedical data is crucial to facilitate the deployment of recent research results based on the specific characteristics of every individual. The vast amount of distributed data marts of diverse, evolvable, and not always reliable data, the fact that current studies in the biomedical field have to do the actions to integrate and process data in a manual way, and the diversity of constraints for data access, are good reasons why the MOSAIC, System, through its support on the achievement of multilateral data exchange agreements, can increase the efficiency of organisations, specifically those devoted to clinical and genetic research. At the end, this efficiency can be measured in terms of the amount of data accessible for a certain study and time used for obtaining the access rights to the data.

3.2 Federated Data Bases and Data Warehouses

Data Base Management Systems is a mature area of computer science and also a mature segment of the IT market where a number of alternatives are available. MySQL, PostgreSQL, Oracle, SQLServer, DB2, Sybase and Teradata are some of the most common Relational DBMS currently used both in the industry and academia. During the last years, a clear tendency to use open source systems more intensively, namely MySQL or PostgreSQL, is changing the market and user preferences. Moreover, the advent of the Big Data era demands novel strategies to manage the Big Data Bases due to their specific characteristics of volume, variety, velocity and veracity. Cassandra [2], MongoDB [80] and HBASE [4] are NoSQL DBMS for the massive storage and computation of big datasets.

1. **Cassandra** is designed for the storage of critical data that is replicated to multiple nodes for fault-tolerance and linear scalability.

2. **MongoDB** is based on the storage of collections of fields and values in a structured form. It uses sharding for horizontal distribution of the data across multiple sites and also distributed replication for critical data storage.

3. **HBASE** supports the distributed data storage of billions of records. It runs on top of the Hadoop Distributed Filesystem (HDFS) that with YARN (a framework for job scheduling and cluster management) and MapReduce (a system for parallel processing of large data sets), compose **Hadoop** [3], a framework for the processing of large distributed datasets (terabytes in thousands of nodes).

In 1985, Dr. Edgar Frank Codd defined the twelve rules that characterise a relational databases [20]. Eight years after that, he proposed a new set of twelve rules to define the Online Analytical Processing (OLAP) [19] for the multidimensional databases or Data Warehouses (DWH) of Decision Support Systems (DSS), which have specific features, are build according a particular architecture, and can be centralised, distributed or federated [96, 114, 9, 76]. DWH are feed by heterogeneous sources of information, build according a multidimensional structure, need periodic but not permanent inserts, and receive several and complex queries to support the data mining. See Fig. 3.2. Updates of the information and transactions in a DWH are unusual.

The interoperability of federated DB and DWH becomes extremely complex when the local databases or Data Marts are heterogeneous and implemented with different schemes. The correct association between fields of different Data Marts needs the understanding of the meaning of each field and the Web Semantic research area [50, 10] is providing powerful tools to face this challenge. The definition of an Ontology is a way to standarise and formalise the semantic representation of the information hosted in a database. The schema of a database can be

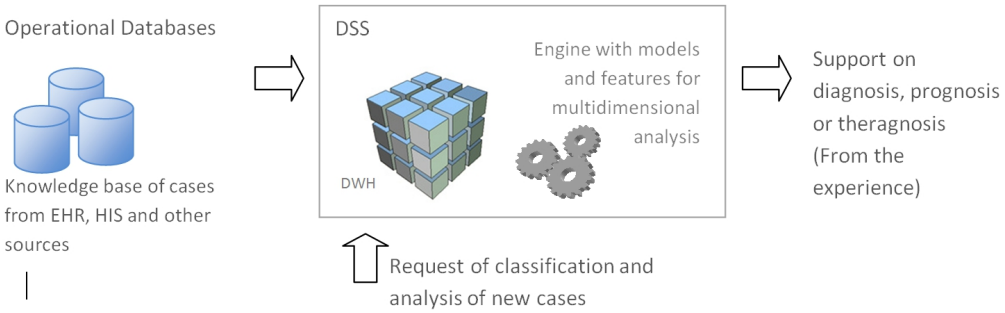

Figure 3.2: Conceptual architecture of a Decision Support System (DSS) with its Data Warehouse (DWH)

mapped with an ontology [94] that represents the domain, and a single ontology may have mappings with a number of different schema, which represent the same concept in different ways. The lack of a uniformly structured data across related biomedical domains is an example of this problem and also an scenario example to show how this problem can be addressed using Semantic Web techniques [95].

The retrieval of information from heterogeneous databases using an Ontology mapped to all of them can be easily done, while the insert and update of a database through the Ontology is a complex task, if not impossible in some cases.

Link Open Data [97] is a W3C Community project that provides a framework to build graphs by setting links [107] between Open Data sets. These links allow to navigate from a data set to those other related data sets through a semantic web browser. Fig. 3.3 shows the Linked Open Data network of 570 data sets connected by 2909 links currently available.

The implementation of an Ontology-based federated DWH with a set of heterogeneous Data Marts is specially attractive as a DSS or a recommender system needs to permanently retrieve data from the DWH, but do not need to update or insert data. In a federated DWH the different Data Marts are likely feed by local systems and the direction of the data comes from the DWH to the DSS and not vice versa. In addition, data sharing and exchange in a network requires the establishment of agreements among a set of nodes that have to negotiate and agree on the terms of the collaboration. The communication and negotiation among a set of networked nodes can benefit from a semantic representation of the terms used during the process. An example of this is Linked USDL [88], an ontology that supports the trading of services over Internet. The most popular tool for the design of Ontologies is Protege, that in its current version 5 supports OWL 2.0.

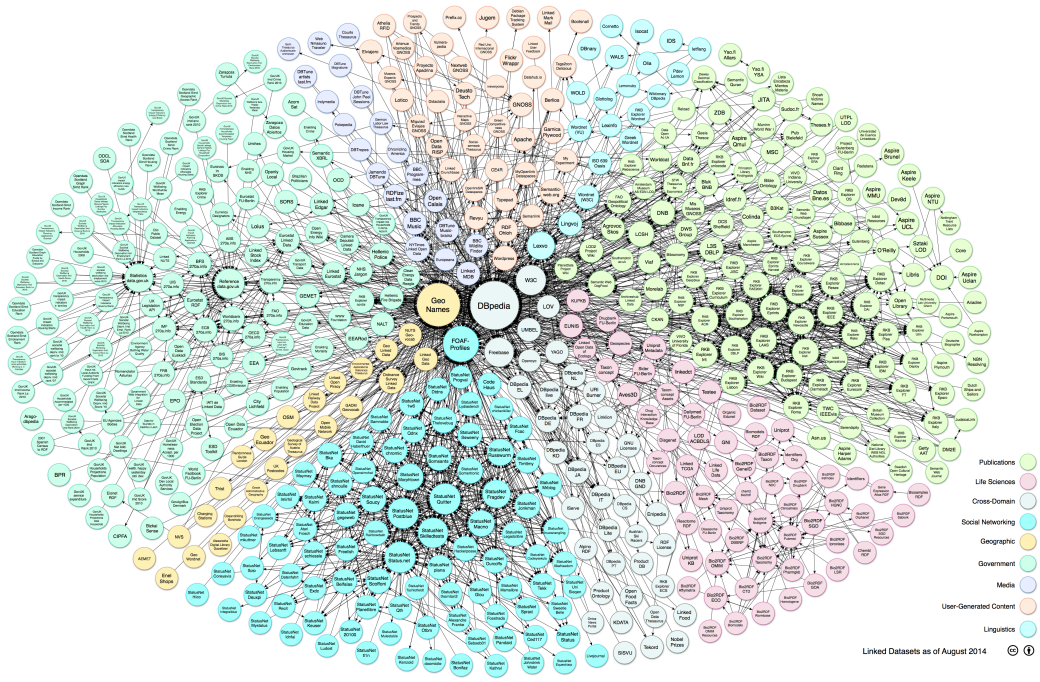

Figure 3.3: Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/.

3.3 Intelligent and Autonomous Communications

In addition to the research and advances in distributed databases, a fertile research area that has greatly contributed to enable the communication and negotiations among "independent" entities of a distributed system is the Agent Technology through Multi Agent Systems.

The specific features that characterise an "Intelligent Agent" [16] are: i) Internal knowledge-based state that can be dynamically altered; ii) Dynamic reasoning capabilities that determine their internal behavior through constraints or goals; iii) Communication status that enables them to interact with other agents or human entities; and iv) Unique identity that provides roaming and service advertising capability. All of this provides flexibility and modularity in developing the infrastructure for an intelligent and dynamic cooperation of federated DB and DWH. Different geographically independent databases and analysis tools can be directly used as a rigorous knowledge base for the agents.

Agent frameworks for medical purposes have been extensively documented in the literature [81, 78, 1, 60, 59]. Although their use started with patient control and healthcare management [53], nowadays there is a wider use in other medical and related domains, including

Security, DSS, Planning, Simulation and Data Management. Examples of this research are FUSION [21], a Multy-Agent System for the integration of applications and services in the area of Ambient Assisting Living, and AIDA (Agency for Integration, Diffusion and Archive of Medical Information) [18], an intelligent agent-based platform to ensure interoperability in healthcare units. There is also a starting research applied to data sharing, as shown first in the recent research work on "Dynamic Health Data Aggregation" [31] presented at the IX Workshop on Agents Applied in Health Care, that already cites some work derived from this Thesis, and gives new approaches to integrate healthcare data using Multy-Agent Systems, focusing on the security aspects of the data integration.

3.4 Standards

Protocols for data transfer and ontologies for the semantic representation of the data are the two main research areas contributing to the understanding and integration of heterogeneous federated DHW hosted in different systems and platforms. The implementation of Intelligent Agents and the communication protocols to support their dialogs follow the specific standards designed at different levels and for specific characteristics. When a multilateral agreement required for the collaboration among a set of nodes is achieved, the data exchange can benefit from the use of standards for both data representation and data transfer.

3.4.1 Data representation

The Systematized Nomenclature of Medicine (SNOMED CT) [101, 100] is the main standard for medical terminology, maintained by the International Health Terminology Standards Development Organisation (IHTSDO) [55]. SNOMED was created in 1999, has more than 300.000 medical concepts, adopted by more than 20 countries and translated in several languages, facilitating its local deployment worldwide. The International Classification of Diseases (ICD-10) [109] provides a detailed list of diseases, symptoms and clinical protocols. ICD-10 was created by the World Health Organisation, adopted by about 25 countries and translated into 42 languages. A part from medical concepts, the other forms of non-text clinical data (e.g. MRI and ECG) also need standards. As an example of this, Digital Imaging and Communications in Medicine (DICOM) [90] is the specific standard that provides support for representing medical images and signals (DICOM-Wave).

The use of these standards facilitate the homogeneous representation of the clinical knowledge in Electronic Health Records (EHR) [63] and allows a common understanding of the clinical data worldwide distributed. Epidemiological studies and combined analysis of a number of clinical data sets benefit from the use of these standards. And in addition to this, also the integration, sharing and exchange of clinical data from federated and distributed DWH.

3.4.2 Data transfer

There is a number of well established interoperability standards for transferring clinical information. Most relevant are HL7 [62] and CEN/ISO 13606 [61].

HL7 defines several groups of standards and implementation guides: HL7 messages (e.g. v2.6 messages with XML syntax; v3.0 messages - interoperability specification for transactions that are derived from the HL7 V3 Foundation models and vocabulary and define communications produced and received by computer systems, they include the concepts of message wrappers, sequential interactions, and model-based message payloads), HL7 CDA (Clinical Document Architecture – for exchanging clinical documents across specialists), HL7 RIM (Reference Information Model), HL7 PHMR (Personal Health Monitoring Report). Part of these standards are approved by ISO and ANSI.

CEN/ISO 13606 is the European norm accepted as international standard for the transmission of the patient's data stored in EHR. This norm facilitates the interoperability of EHR repositories and distinguishes two categories of data, namely information and knowledge. The first one corresponds to the Reference Model and standardises the representation of information. The second one corresponds to the Archetypes providing a formal representation of clinical concepts through a structured combination of the information represented by the Reference Model.

Besides HL7 and CEN/ISO 13606, the DICOM standard, in addition to the formal standardisation for data representation, provides also a specific protocol for transferring DICOM objects and specifies their encoding using the HL7 architecture, allowing the interoperability between DICOM and HL7.

3.4.3 Intelligent Communication

The standardisation of agent communication, agent transport, agent management, abstract architecture and applications platforms, was promoted by the Foundation for Intelligent Physical Agents (FIPA) [40], an organisation that started in 1996 and although during recent years has been almost inactive, was accepted as the IEEE eleventh standards committee on 2005 and is still alive. The IEEE-FIPA standard specifications are grouped by i) The Agent Communication Language (ACL) [42], which specifies the format of the messages to be used by the agents for their communications, message exchange interaction protocols, speech act theory-based communicative acts and content language representations; ii) The Agent Management specifications that deal with the control and management of agents within and across agent platforms; iii) The Agent Message Transport specifications, that deal with the transport and representation of messages across different network transport protocols; iv) The Abstract Architecture specifications, that deal with the abstract entities that are required to build agent services and an agent environment; and v) The application specifications set, as example application areas in which FIPA agents can be deployed.

- ACL includes: i) eleven specifications of Interaction Protocols that deal with pre-agreed message exchange protocols for ACL messages; ii) the Communicative Act (CAs) specification, that deal with different utterances for ACL messages; and iii) four Content Language (CL) Specifications that deal with different representations of the content of ALC messages.
- Agent Management integrates three specifications
- Agent Message Transport is composed by i) three Message Representation specifications dealing with different representation forms for ACL messages; ii) two Envelope Representation specifications, that deal with different representation forms for ACL message envelopes; and iii) three Agent Message Transport Protocol (MTP) specifications, that deal with different network transport protocols (IIOP, WAP and HTTP) for delivering ACL messages.

Although the FIPA specifications of the Agent architecture are mature in the indicated areas, and their acceptance as standards is generalised, their implementation to create the infrastructure needed for the agent system development is an ongoing task [45]. The most popular Agent Platforms, namely FIPA-OS, JACK, ZEUS and JADE [65, 7], among others, are FIPA compliant. Ideally, agents could interact with agents written in other languages and running on other platforms. In JADE, like in some other agent platforms, messages among the agents are written using ACL, granting the understanding with agents running in other platforms also ACL compliant.

The deployment of Intelligent Agents into a specific domain needs a specific Application Specification, including the ontology of the area and the service description. A number of Application specifications are designed to work with FIPA standards. These are the following: a) Nomadic Application Support Specification, b) Agent Software Integration Specification, c) Personal Travel Assistance Specification, d) Audio-Visual Entertainment and Broadcasting Specification, e) Network Management and Provisioning Specification, f) Personal Assistant Specification, g) Message Buffering Service Specification and h) Quality of Service Specification.

The life cycle of a FIPA standard has five states and only two of the current eight specifications at Application level have the mark of accepted standard. The majority of them, either standards or standard proposals, are defined according to the following structure: i) The scope and a general analysis of the specification, ii) The scenario of the Application; ii) The Ontology associated to the specification, with its 'object descriptions', 'function and predicate descriptions', 'interaction protocols' and 'exceptions'; and finally iii) Some Examples and References.

At the moment, there is not any FIPA specification, at any stage, for clinical data sharing among agents.

Strong security based methods are essential in many areas, like health. In the agent communications based on Jade, the security can be provided by the JADE-S midelware [106].

Although at present there is not any formal extension of UML to design Multi-Agent Systems with all their specific characteristics, the overall design of communication protocols among agents [77] can be designed using standard UML Sequence Diagrams or specific designing tools for MAS (e.g. Wade and Wolf [39]).

3.5 Related projects

Based on the 2014 roadmap of the European Strategy Forum on Research Infrastructures (ESFRI [35]), Europe is making efforts towards a competitive infrastructure for bioinformatics (BBMRI-ERIC [37], ECRIN [33] and Elixir [36]). Big challenges have to be addressed and this is why EU included "omics in personalised medicine" as part of its goals for the EU 2020 strategy, where interoperability and data aggregation are hot topics (e.g the NHGRI 2012 Workshop [102] and the Global Alliance initiative [49]).

Different on-going projects work towards the integration of distributed and heterogeneous data sets at the individual and population levels (e.g. Gen2Phen [43, 108], and Ensembl [34, 41]) and several projects in different domains are addressing issues related to the integration, sharing, access and analysis of large amount of data (e.g. GeoKnow [44, 113, 5], Linking Open Data [97], IQmulus [58], SemaGrow [98], Insight [57], LDBC [67, 13], IMPART [56], BIOPOOL [12], Optique [86, 46], Biobanckcloud [11] and [6]). Nevertheless, none of these projects study how to facilitate multilateral exchange of the huge, diverse and disperse amount of distributed data available worldwide. MOSAIC will provide a novel, flexible and scalable approach to data access in distributed, large and heterogeneous sources.

Aiming to extract the maximum amount of knowledge from the data, a global alliance [32] for sharing genomic and clinical data was created on june 2013.

Cancer Biomedical Informatics Grid (caBIG) was one of the biggest initiatives working to provide an infrastructure to build distributed databases, to share data and knowledge. "The Cancer Genome Atlas" (TCGA) [82] was one of the examples where a distributed database was built using caBIG for its development. The required semantic interoperability in caBIG was provided by caCORE [66] service oriented architecture.

The Artemis Project [14] has previously dealt in a medical context with the transfer of critical patient data (mainly clinical data). Artemis used a peer-to-peer-like architecture and web services security protocols for the transfer of patient's records.

HOPE [52] is a collaborative telemedicine platform for clinical data sharing. It implements a grid infrastructure for a distributed database, providing a common interface independent to the data base of its nodes, to manage data from the patients both textual and graphical, accessing to clinical records and images (in DICOM) from PACS.

An example of a federated Data Warehouse and its associated Decision Support System is

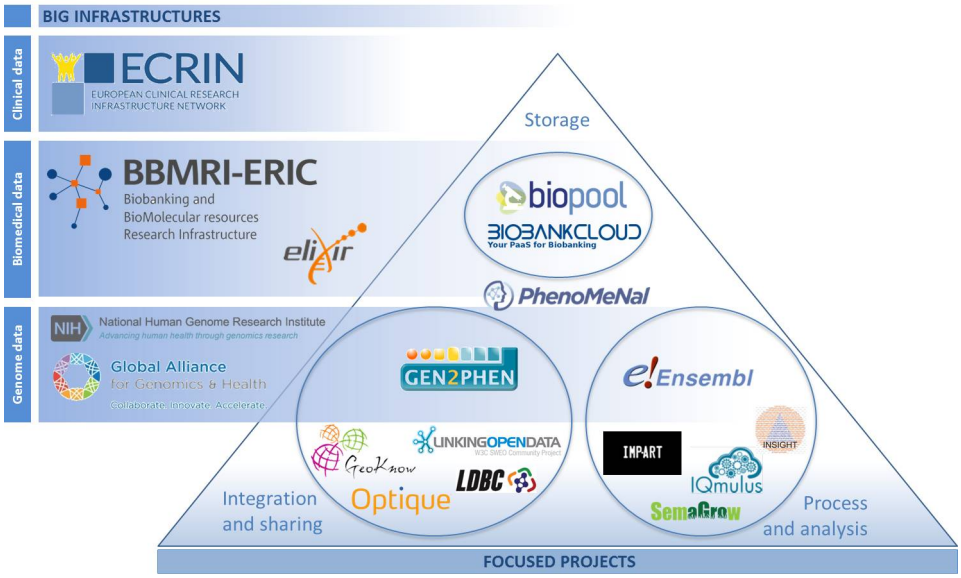

Figure 3.4: Main infrastructures and a sample of relevant projects aiming to facilitate data integration, sharing and analysis, both from the biomedical area and from some other fields.

the HealthAgents project [73], that aims to build a worldwide network of clinical centres for the brain tumour diagnosis.

More recently, Phenomenal [89] is a project to create an e-infrastructure to process, analyse and mine medical molecular phenotyping and genotyping data from clinical and population studies.

All these systems (see Fig. 3.4 for a combined view of all of them) implement and provide either a good framework for distributed clinical databases where to store the data in a virtual repository or federated databases where the retrieval of the data among the nodes is performed after a direct query to the system. Nevertheless, at the moment, there is not any system that performs an automatic agent-based negotiation for clinical data exchange in a federated DWH.

The MOSAIC System, provides the functionalities and features that allow to perform an automatic negotiation for the clinical data exchange, by using a Multi-Agent System.

Part III

Contributions Towards Intelligent Clinical Data Exchange

Chapter 4

The Architecture of MOSAIC

"Perfection is achieved not when there is nothing more to add, but rather when there is nothing more to take away"
Antoine de Saint-Exupery (1900-1944)

4.1 The overall picture of MOSAIC

4.1.1 Main components of the system

The MOSAIC System (see Fig. 4.1) is composed by a network of interconnected nodes each one with its associated Data Mart and on top of this a Multi-Agent System (MAS) to manage the communication among the nodes and negotiate the data exchange among them. The agent oriented abstraction fits well in this knowledge sharing scenario due to its distributed and dynamic nature. The MOSAIC MAS facilitates the multilateral data exchange in the network by providing mechanisms for the intelligent search of paths to reach the datasets requested, involving a subset of the network nodes in multilateral agreements.

The MOSAIC Data Marts compose all together a Federated Data Warehouse (DWH). The Data Base Management System (DBMS) of each Data Mart can be implemented by different solutions, either traditional systems like MySQL, ORACLE or Postgres, or novel DBMS like Casandra, MongoDB or HBASE. In all cases the data hosted in each DataMart must have the minimum common data representation to allow the interoperability and build the common framework of the Federated DWH. To this end, each Data Mart will have a common format to reference the data types hosted, and all meta data and metrics required for the negotiation process, including the size of the datasets hosted, the constraints to allow its access, and if applies, the reference of the requested data (as constraint to be solved).

The specific clinical data hosted may have different formats, but in order to facilitate the data exchange when the agreements are achieved, data format should follow the current clinical standards for data representation (e.g. DICOM, SNOMED, ICD-10, etc) and each Data Mart

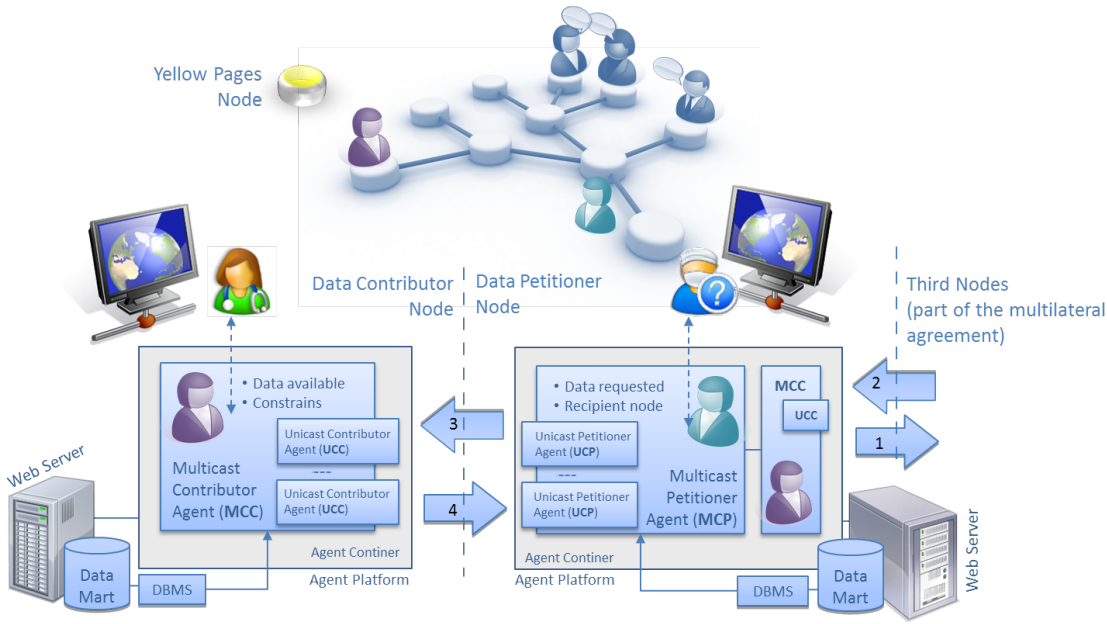

Figure 4.1: The architecture of the MOSAIC System, showing the data flow among the agents. The Data Petitioner Node solves the constraint: 1) data delivery to thrid nodes, 2) data collection from third nodes needed to fulfil a constraint and 3) delivery of data requested. The Data Contributor node concludes the transaction: 4) delivery of the data requested.

should also have in place the clinical protocols for data exchange (e.g. HL7 and EN13606). Every implementation of the MOSAIC DWH will define the common set of data representation and data transfer standards and protocols to make possible the integration of its Data Marts.

Besides the Data Mart and in addition to the issues related to the data storage and transfer, each node of the MOSAIC network has an Agent Container Platform to host and manage its Agents. Although Jade is the most popular Agent platform, similarly to the heterogeneity of DBMS, the Agent Containers in each node can be implemented by different solutions (ZEUS, SkeletonAgent, EVE, MASON, etc). What is most relevant is that the Agent dialog is performed following the same communication language with the use a common standard (ACL FIPA compliant).

Finally, each Node has to provide the interface to communicate with the user. MOSAIC proposes to facilitate this communication through a browser and a Web server (e.g. Apache) is a required component to be installed in each Node.

In summary, a Node of the MOSAIC System is composed by three main components, for i) Data Management, ii) Agent Communication, and iii) User interaction.

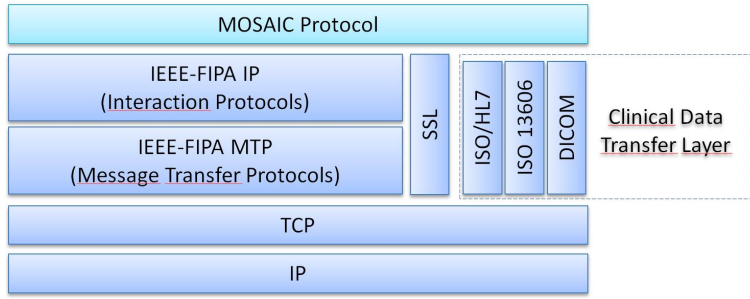

Figure 4.2: The Stack of the main protocols where MOSAIC sits

4.1.2 The Nodes of the MOSAIC network

The nodes of the MOSAIC network have two different roles: One as "*data contributor*" and the other one as "*data collector*". Each Node can play the two roles, only a single one, or it may also happen that a Node is temporary inactive.

A Node becomes *data contributor* when its user decides to share one or more datasets, and *data collector* when the user launches a request for data access. Linked to each of these actions for data request or data sharing, the corresponding Agents will be activated, putting the network aware of the new state of the Node.

During the negotiation process the Agents may require to interact with the user (to accept or reject constraints or to define them when a data access request arrives).

The result of the actions and negotiations of the Agents conclude with new data collected and data delivered. The new data collected will be hosted at the Data Mart of the Node with the reference of the Node where it comes from. For the data delivered the Node will keep record of the recipient Nodes, allowing the user to keep control of the Data Mart, and manage its role in the MOSAIC System.

4.2 The MOSAIC Protocol

4.2.1 The MOSAIC position in the stack of protocols

The MOSAIC Protocol allows the message and information exchange needed during the negotiation process of the MOSAIC MAS aiming to achieve agreements for the data exchange and it is located within the current protocols' stack of the Agent systems and clinical data transfer (see Fig. 4.2).

When a negotiation agreement is achieved and the data transfer has to take place, the MOSAIC Protocol will take advantage of the clinical data protocols. The information hosted

Table 4.1: Agents in the MOSAIC System

ID	Agent Name	Agent Description
MCP	Multicast Petitioner Agent	Agent activated by the user or by a Unicast Petitioner Agent. The user launches it in order to explore the network looking for a certain data set. The UCP launches it in order to solve a constraint from a UCC when the dataset requested is not available at the node of the UCP
UCP	Unicast Petitioner Agent	Agent activated by the MCP in order to negotiate a specific data access request with a UCC
MCC	Multicast Contributor Agent	Agent activated by the user to offer a certain dataset to the network, with or without constraints
UCC	Unicast Contributor Agent	Agent activated by the MCC to negotiate a specific data access request sent by an MCP
YP	Yellow Pages	Agent that provides the directory service and hosts the list of references of MCCs active in the network

at the Data Mart can be part of the Electronic Health Record of the Node's platform, and HL7 and EN13606 will be used for a standard retrieval and delivery of the information. In each link of a multilateral agreement the two participating nodes will start a communication using some of those standards, transferring the corresponding dataset that can include textual or multimedia information. The representation format used and agreed within the MOSAIC System (e.g. DICOM, ICD10 and SNOMED-CT) will facilitate the access and use of the data exchanged by the recipient Node. For security purposes, this data transfer will also use the SSL protocol to guarantee the delivery of the data to the right Node and providing a secure channel for the data transfer during the communication.

4.2.2 The MOSAIC Agents, the main actors of the protocol

Two main different types of Agents are defined in order to act on behalf of a node in the network: These are the "Petitioner Agents" and the "Contributor Agents". Each node may have a set of these Agents running concurrently. Besides, in the network there are other Agents called "Yellow Pages Agents" responsible to maintain the information of the network topology, with the list of the active nodes and agents. The Petitioner Agents will ask the nearest Yellow Pages to obtain the list of nodes to whom they may address their data access requests. Table 4.1 summarises the five types of Agents of the MOSAIC MAS, and Fig. 4.3 shows the dependences and relationships between them.

Agents for requesting data

The Multicast Petitioner Agents (MCP) are launched by the user of a node with a request to collect a certain set of data from the network. They are responsible to identify which are the nodes that may contain the requested data, negotiate with them the access rights, try to solve the conditions and constrains (if any) and finally, collect the data, if possible. The MCP

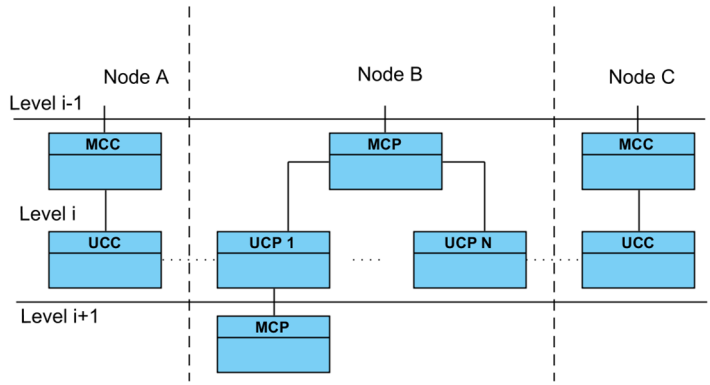

Figure 4.3: The MOSAIC Agents, their dependency and interactions during the execution of the protocol. Every link in a multilateral agreement is composed of two pairs of MCP-UCP and MCC-UCC. YP Agents are not shown, but they interact with MCC and MCP.

activates a Unicast Petitioner Agent (UCP) for each contributing Node offering the requested data. As there may be a number of contributing nodes, candidates to establish an exchange agreement, the MCP launches a UCP per each of them in order to explore and negotiate each possible agreement.

Agents for delivering data

The Multicast Contributor Agents (MCC) are launched by the user of a node that wants to share certain data hosted in its local Data Mart. The data might be made openly available, but also it may have certain constraints to allow its delivery. These Agents will negotiate (if necessary) the conditions to provide access to the data with the Petitioner Agents, and it will deliver the data when the conditions are fulfilled. When the MCC receives a request from a MCP, it launches a Unicast Contributor Agent (UCC) to explore this tentative collaboration.

Agents for coordinating the process

The Yellow Pages Agent is to be launched in certain nodes that will provide the mandatory support to all the Agents of the network to maintain the minimum information needed to establish the first contact among the Nodes and their Agents. Its main purpose is to provide the MCP Agents with the list of active Nodes and their MCC to whom they may try to request the access to their Data Marts. New nodes and new Agents must register to a Yellow Pages Agent in order to be part of the Network. The Yellow Pages Agent is responsible to maintain the list of active nodes (IP addresses and IDs) and their active Agents (also identified with unique IDs). All the Yellow Pages Agents active in the network must be periodically synchronised among them, trying to have updated mirrors of their lists. The protocol for the "Yellow Pages"

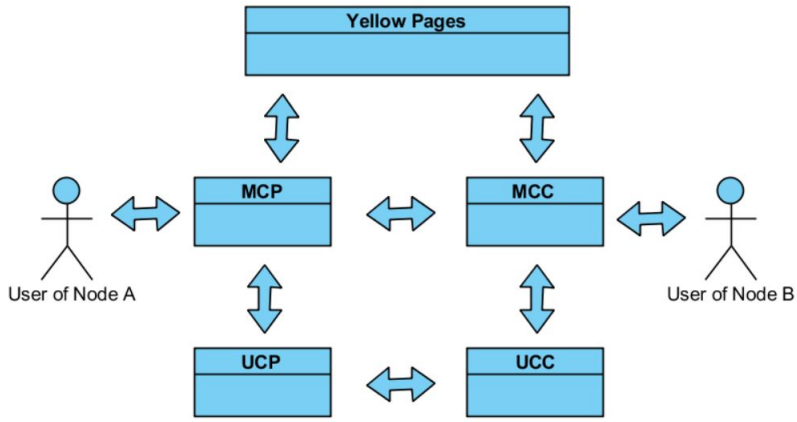

Figure 4.4: The dialog and interaction level of the MOSAIC Agents. There is no communication between the MOSAIC actors not linked with some narrow.

maintenance and update, is out of the scope of this work, and could be implemented like the well know Internet service of the "Domain Name Servers" (DNS).

4.2.3 The communication strategy of the MOSAIC Agents

The interaction between the MOSAIC Agents respects the following principles: i) The users of the protocol interact with MCC and MCP, ii) Unicast Agents are created by Multicast Agents to negotiate every possible data exchange between two nodes, iii) MCP interact with MCC, YP and UCP; and iv) MCC interact with MCP, YP and UCC; and iv) Direct communications between MCP and UCC or between MCC and UCP are avoided. See Fig. 4.4

The MOSAIC Protocol can be parametrised with a set of timers, counters and flags, to manage the state of the negotiation process and the transfer of data among them. This will allow different configurations and uses of the protocol. This flexibility allows the use of the protocol in different scenarios and adapt the communication among the nodes to the behaviour of the Agents.

Chapter 5

The MOSAIC Protocol

"Viam qui nescit qua deveniat ad mare, eum oportet amnem quaerere comitem sibi"
(The one that doesn't know the way to the sea, should seek a river for his companion)
Plautus (254-184 BC)

5.1 A collaborative process for data exchange

The communication and negotiation process of the MOSAIC Agents is composed by five stages: i) Publication of the datasets available from contributor users, ii) Launch of dataset requests by petitioner users, transaction start, network exploration and agreement negotiation, iii) Agreement selection, iv) Delivery of the data, and v) Transaction completion. The main actions involved in each of these stages are summarised in Table 5.1 and described bellow.

Stage 1: Dataset publication

A user activates a MCC and notifies the YP about its existence and the availability of a certain dataset. The dataset can be made public with no constraints or its delivery restricted to the exchange of some other dataset and the acceptance of certain constraints, but this will be part of the second stage of the process and won't be published at the YP, that will keep only the references of the MCC and the associated dataset. The MCC remains active until the user stops it. When this happens, its reference in the YP is removed.

Table 5.1: Stages of the process to build multilateral agreements

Stage		**Main actions involved**
1	Dataset publication	Publication of the dataset available
2	Network exploration	Start of transaction, dataset request, and build of the agreement path
3	Agreement selection	Transmission of the authorisation and access rights to the data
4	Data transfer	Delivery of the data and Acknowledge the reception of the data
5	Transaction completion	Commit or Rollback the whole transaction

Stage 2: Network Exploration

After the activation of a MCP by a user, the process to find paths that connect the requesting node with the ones hosting the desired data starts.

In this stage the MCP asks the YP to obtain the list of MCC to whom the data access requests can be addressed. The reference of the MCC delivered by the YP are those hosting a dataset of the type requested by the MCP. The process of the agreement exploration seeks for paths composed by a set of nodes connecting MCC offering the requested data with the MCP (directly or with intermediate connections with other nodes) and concludes with: i) a successful result, providing possible multilateral agreements for data exchange or ii) a failed result, without any path connecting the MCP with some MCC candidate.

Stage 3: Agreement Selection

Every agent participating in a successful path will notify its creator about the possible agreement. At the end of this stage the initial MCP will receive a list of all existing possible agreements for the data exchange (corresponding to a list of paths that go from the leaf to the initiating MCP).

The MCP will select a path or a set of paths and notify this decision to all the agents involved, considering - among other criteria - to avoid overlapping agreements that solve the access to the same dataset of the same MCC through different paths, or to datasets already collected.

Stage 4: Data Transfer

After receiving the notification that a possible agreement is selected, the data exchange between all the nodes starts. This may end with a complete and successful data exchange or with some failure by some nodes. All the UCP waiting to receive data will send a message of acknowledgement (ACK) to their MCP after receiving the data or a message of non-acknowledgement (NACK) in case of a failure of data reception. The ACK (or NACK) is transmitted link to link until arriving to the main MCP at the top of the path.

Stage 5: Transaction Completion

After receiving all ACK from all nodes involved in the agreement, the initiating MCP will send a COMMIT to all Agents. In case some ACK is not received or a NACK is transmitted by some Agent, the MCP will send a ROLLBACK message to all nodes. Only after the reception of a COMMIT the nodes will have the authorisation to use the data received. In case the transaction is aborted with a ROLLBACK, none of the nodes of an agreement that received data are authorised to use it.

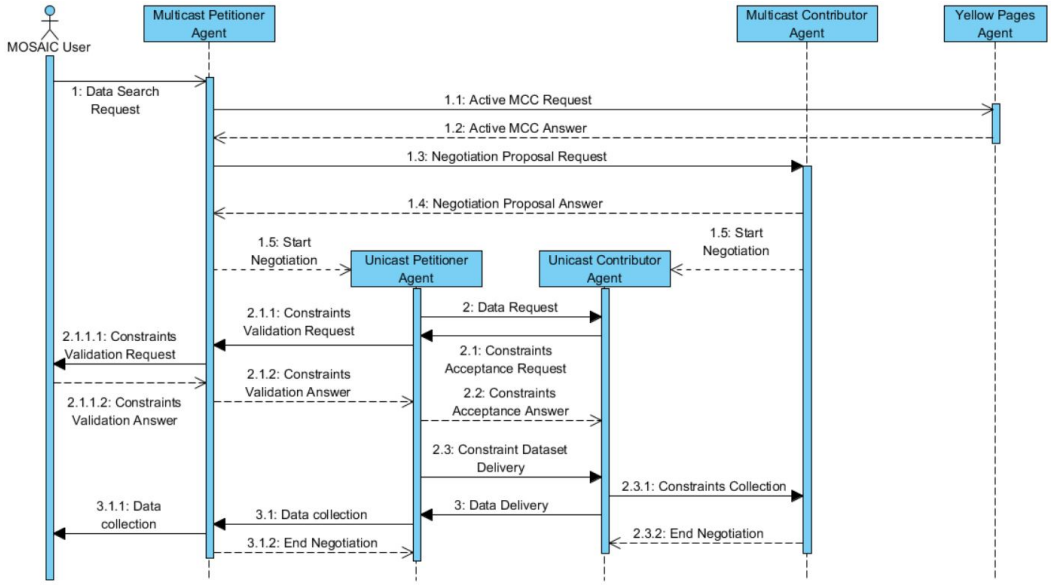

Figure 5.1: Sequence diagram showing the MOSAIC communication protocol used by the MOSAIC agents when a data search request is launched

The dialog and communication among the MOSAIC Agents (MCC, MCP, UCC, UCP and YP) required to perform the negotiation process for multilateral agreements after a request is launched, is represented in the sequence diagram shown in Figure 5.1.

5.1.1 Contributing and delivering data

The user responsible of the DataMart in a certain node launches the MCC when wants to put a Dataset available in the network. The user also indicates to the MCC which are the constraints to be fulfilled before allowing the access to the data. After its activation, the MCC publishes to the Yellow Pages its existence and waits to receive requests from the active MCPs.

When a new request arrives, the MCC launches a new Agent (UCC) to process it with the goal of validating the possible data access constraints and delivering the data if the constraints are fulfilled. If the data is published with no constraints, the access will be automatically accepted and the data delivered. Otherwise, the conditions and constraints for accessing the data are sent to the MCP and the MCC waits to receive the answer, accepting or rejecting the conditions and checking whether the constraints are fulfilled or not. It might happen that the MCP agrees partially with the conditions and proposes some new agreement to the MCC who will ask to the user whether the new proposed agreement may be accepted or not. A timer and a counter will put a limit to the time for achieving the agreement and the number of interactions that the MCP and the MCC may explore for the agreement.

5.1.2 Requesting and collecting data

After receiving the request from the user with the details of the data set to be collected from the network, the MCP consults the Yellow Pages which are the active nodes of the network with MCC running, and sends to all of them a message with the details of the data set that aims to collect and the authorization access request to those that have data that may be part of that data set. Then, the MCP waits for receiving messages from the MCCs indicating the existence of the requested data that may be accessed either with or without conditions.

When a set of options (MCC candidates) are found, a set of new UCP (one per option) will be launched in parallel and autonomously will explore each of the possible options for data collection. Each of these UCP will ask to the corresponding UCC which are the conditions for accessing the data, if any.

For the UCC that do not have any constraint and offer their data openly, the process for collecting the data available will be launched directly. In case there are some conditions to be accepted prior the delivery of the data, the MCP asks to the user whether those conditions are accepted or not. If the conditions are fulfilled and accepted, the UCP will proceed with the data collection.

When the condition for accessing to a certain data set of a node is to deliver another data set that the node might be interested to have, two situations may arise:

- **Bilateral data exchange**
  UCC's condition for allowing access to its data set is to receive another data set available at the Data Mart of the UCP. This agreement depends only to the bilateral data exchange between the Contributor and the Petitioner Nodes. The UCP asks to its MCP to look for the MCC active in its node in order to collect the data from its DataMart. The UCP sends the notification of the dataset availability at the UCC after the potential fulfilment of the constraint. The UCC sends the agreement for the possible dataset transfer initially requested to the UCP. Both UCC and UCP notify their agreement for the potential exchange of the corresponding datasets to their MCC and MCP.

- **Multilateral data exchange**
  UCC's condition for allowing access to its data set is to receive another data set, not available at the Data Mart of the UCP. The constraint of the UCC cannot be resolved locally and forces the UCP to look for the requested data set in other nodes of the network.

  The access to the data set of the UCC depends on multilateral agreements among a set of nodes in the network. A "time to live" parameter (TTL) will be defined to allow the user of the protocol to set a limit at the number of nodes that may be involved in a multilateral agreement. This will allow to avoid unmanageable network explorations. If the length of the path does not exceed the TTL limit, the UCP launches a new MCP to look for the

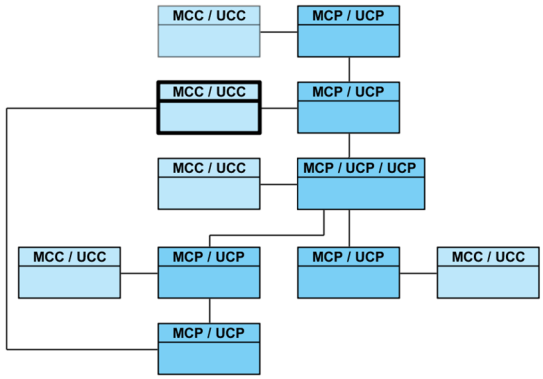

Figure 5.2: Example of a path in the network exploration where a MCC (in bold) will deliver its dataset without solving its constraint after identifying a loop.

data set needed in order to solve the constraint and starts a recursive process.

A MCP activated by another Agent in order to solve a constraint has to take into account that the final recipient of the requested data set would not be the node from where the Agent is launched and this has to be notified to the possible contributing nodes. Moreover, the data access request does not have to be addressed to all the active Contributor Agents as it has to exclude the recipient node of the requested data.

When the access to the data cannot be solved, the MCP will wait a certain time to allow a reconfiguration of the network and its content. After this time, the MCP will start a new request. This may happen a number of times until a counter that puts a limit to this interaction expires.

A node and an MCC can take part more than once in a path of a multilateral agreement, however a special case occurs when in order to solve a constraint of an MCC the subsequent activations of new MCP results in a new request to the same MCC. If the request comes from an MCP "child" (belonging to the same branch), the MCC decides to activate the UCC without any constraint and thus, deliver its dataset without receiving any dataset in advance (see Figure 5.2). After completing the delivery of the other datasets in the path links the MCC receives the dataset of its constraint from the first MCP of the branch that initiated the negotiations.

The exploration of the network may cover all the possible paths (flooding) or a selection of them. The use of flooding is not only inefficient, but in networks of certain size not feasible due to the computational costs. Therefore, when an MCP receives the set of MCC candidates, it will select a subset of them to continue the network exploration.

When an MCP does not find any MCC with the data set needed to fulfill a constraint, it stops the exploration and notifies to its creator UCP on the failure of the path in its attempt to find a multilateral agreement.

5.2 The states of the MOSAIC Agents

5.2.1 The Contributor Agents

The behaviour and life cycle of the MCC and UCC are shown in the state diagrams of figure 5.3 and explained below.

- **C1: Start Contribution**
  This is the first state of the MCC. It receives - as input from the user - the reference of the dataset available and the access conditions to be fulfilled. It publishes to the YP its existence and jumps to the next state "Waiting for Request" (C2).

- **C2: Waiting for Request**
  As its name indicates, the MCC stays in this state attending possible Data access requests from the network. When a new request arrives the Agent launches a new UCC to process the specific request with the goal of validating the possible data access constraints and delivering the data if the constraints are fulfilled. The MCC receives the results from each UCC and notifies the user periodically with those results. A stop request from the user implies to jump to the final state "Closing MCC" (C3). However, before stopping the Agent, it will wait for some time to allow the possible active UCC to conclude. After expiring that timer, in case there is still some active UCC, they will be forced to stop before leaving this state.

- **C3: Closing MCC**
  This final state of the Contributor Agent is to notify the Yellow Pages about its end.

- **UC1: Start Delivery**
  This is the first state of the UCC. It receives the reference of the dataset to be delivered and the reference of the final recipient node where the data will be transfered, including also the reference of the UCP to which the negotiations will start. If the delivery concludes correctly, the Agent will move to the final state "Delivery successful" (UC2). If for some reason the delivery of the dataset can not be concluded, the Agent will move to the final state "Delivery failed" (UC3).

- **UC2: Delivery successful**
  This is the final state of the UCC when the data delivery concludes successfully. A message notifying the achievement of the delivery to the UPC is sent and the Agent ends.

- **UC3: Delivery failed**
  This is the final state of the UCC when for some reason the data delivery can not conclude successfully. A message notifying the failure of the data delivery to the UCP is sent and the Agent ends.

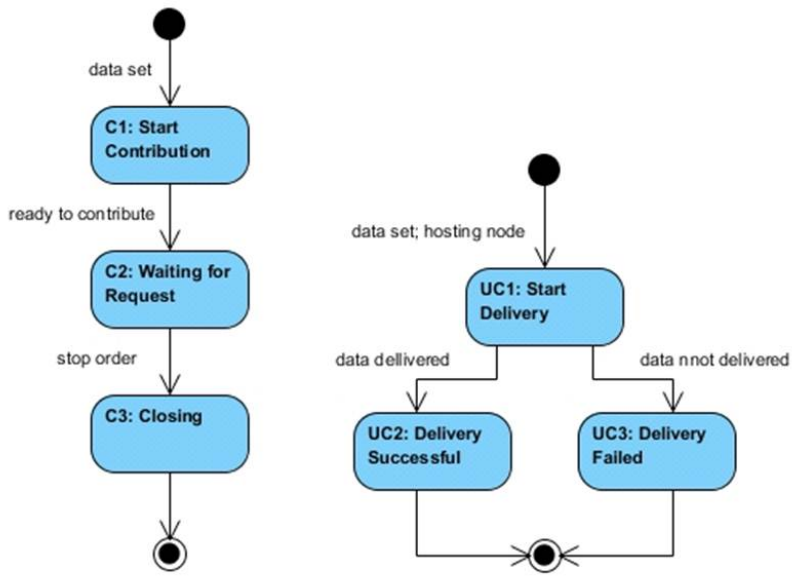

Figure 5.3: State Diagrams of the Multicast (left) and Unicast (right) Contributor Agents

5.2.2 The Petitioner Agents

The behaviour and life cycle of the MCP and UCP is described with the State Diagram shown in figure 5.4 and explained below.

- **P1: Start Request**
  This is the first state of the MCP Agent. It receives the request from the user with the details of the data set to be collected from the network. In this state the Agent consults the YP which are the active nodes of the network with MCC Agents running and sends to all of them a message with the details of the data set that aims to collect and the authorization access request to those that have data that may be part of that dataset. After sending that request, the Agent jumps to state P2 ("Waiting for Option").

- **P2: Waiting for Option**
  This is the second state of the MCP Agent. The Agent waits for receiving messages from MCC Agents indicating the existence of the requested data that may be accessed either with or without conditions. A timer (tWO) indicates the time frame while the Agent will be waiting in this state. After expiring the time indicated the Agent will jump to state P3 ("Option Available") if there is at less one answer from some MCC Agent or to P6 (No Option Available) if there is no answer from any MCC Agent.

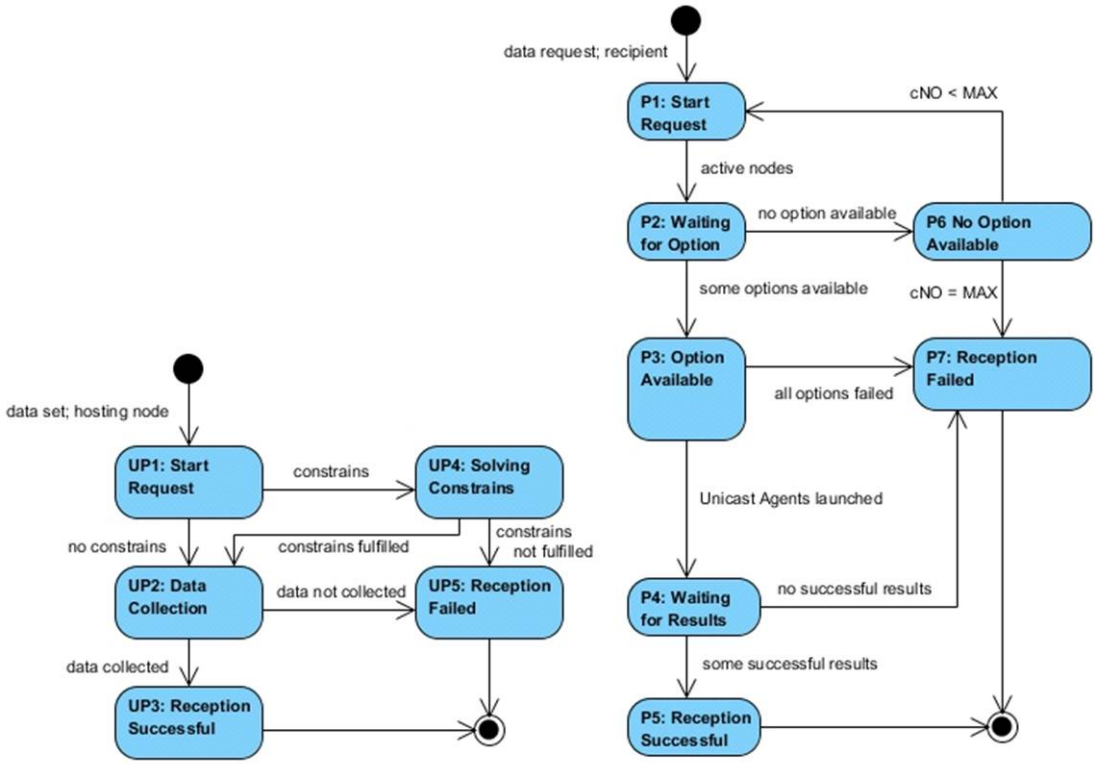

Figure 5.4: State Diagrams of the UCP (left) and MCP (right)

- **P3: Option Available**
  When a set of options are found a set of new UCP Agents (one Agent per option) will be launched in parallel and autonomously will explore each of the possible options for data collection. Each of these UCP Agents will start with direct requests to explore the possible data access to a specific node. After launching the set of UCP Agents the MCP Agent will jump to the state "Waiting for Results" (P4).

- **P4: Waiting for results**
  This is the state where the MCP waits for the results of the UCP launched in the previous state. A new timer tWR will determine the time limit while the Agent will be waiting for the results of the UCP Agents. When the tWR expires or all the results from the new unicast Agents are received, the Agent will examine the results for its final action. In case there is not any successful result from any UCP Agent, the main Agent will jump to state "Reception Failed" (P7). Otherwise, the Agent will jump to state "Reception Successful" (P5).

- **P5: Reception Successful**
  This is the final state of the MCP Agent for successful data requests. In this state the MCP notifies the user about the success of the request and ends.

- **P6: No Option Available**
  In this state the Agent waits a certain time to allow a reconfiguration of the network and its content. This time is defined by timer tNOA. After this time, the Agent will jump back to state P1 "start request". This may happen a number of times until a counter (cNOA) that puts a limit to this interaction does not arrive to zero. cNOA counter is initialised with a value that marks the maximum number of interactions and is decreased with a unit each time the Agent falls in this state.

- **P7: Reception Failed**
  This is the final state when there was not any node with the data requested or with conditions that made possible to achieve an agreement for allowing the access to the data. The MCP Agent fails in its goal to access to the data and ends.

The UCP Agent is automatically activated by a MCP Agent in order to manage an active option for accessing to a specific data-set. It receives the request for accessing to the data set, with the reference of the node where it is hosted, plus the reference of the final recipient of the data. It will negotiate with the UCC Agent from that node the access rights to that data set and eventually collect the requested data if the constraints are fulfilled.

The conditions that the UCP Agent have to fulfill may include i) Specific conditions defined by the owner of the data to be validated by the user of the Petitioner Agent; and ii) Data exchange, either available at the Data Mart of the Petitioner Agents' node or not. It may also happen that the dataset is freely available with no constraints to be satisfied or data to deliver.

- **UP1: Start Request**
  This is the first state of the UCP Agent where it receives the reference of the dataset to be collected and the reference of the node where it is hosted, the reference of the MCC Agent with which the negotiations will start and the reference of the final recipient of the requested data (which may be either the requesting node, or a third node if the request is only to solve a constraint for accessing to another data set). If there are no constraints to fulfill, the Agent will move to state "Data Collection" (UP2). If there are some constraints to fulfill, the Agent will move to state "Solving Constrains" (UP4).

- **UP2: Data Collection**
  In this state, all the conditions, if any, have been already fulfilled and the UCP Agent is authorised to collect the dataset. The transfer of the data takes place here and if it concludes successfully the Agent moves to the "Reception Successful" state (UP3). If

some problem occurs and the data transfer can not conclude, the Agent moves to the "Reception failed" state (UP5).

- **UP3: Reception successful**
  This is the final state of the UCP Agent when the data collection concludes successfully. A message notifying the achievement of the data request to the MCP Agent is sent and the Agent ends.

- **UP4: Solving Constrains**
  This is the state where the Agent comes when there are constraints to fulfill. If specific conditions need to be accepted by the requesting user, the Agent will send a validation request to him through the MCP.

  In case there is a condition to provide another dataset, the Agent will check whether it can be obtained from its Data Mart. If so, an authorisation for delivering it to the Contributor Agent will be requested to the user. If the data can not be obtained locally, a new MCP Agent will be launched. A MCP Agent must know whether its activation corresponds to a need to solve a constraint or not. If so, it must avoid sending a request to a possible MCC Agent active at the node of the final recipient of the data-set. When the dataset is available the data delivery to the UCC will take place.

  Answers from the users and if needed from the new MCP Agent and / or from the UCC, are managed in this state. In case all the constraints are fulfilled, the Agent will move to "Data collection" state (UP2). If some constraint can not be solved, the Agent will move to "Reception failed" state (UP5).

- **UP5: Reception failed**
  This is the final state of the UCP Agent when for some reason the data collection can not conclude successfully. A message notifying the failure of the data request to the MCP Agent is sent and the Agent ends.

5.3 The MOSAIC implementation

The MCP has been implemented according to Algorithm 1, the MCC is presented in Algorithm 2, the UCC in Algorithm 3, and the UCP implementation in Algorithm 4. In order to clarify the process and to highlight only the most important features of the protocol, the pseudocode presented here merges the steps of stages 2 to 5 of the negotiation process and after an agreement, the dataset is directly transferred to the requesting agent.

Two important aspects of the implementation correspond to i) the way that a path of a possible agreement is created and propagated and ii) the way a loop is detected.

Algorithm 1 Multicast Petitioner Agent *(MCP)*

Inputs

ResourceRequested from *User* or UCP'
MCC from YP
ResourceDataset from UCP
NegotiationAgreement from UCP

```
1: Ask YP for MCC hosting the ResourceRequested
2: Collect MCC compatible from YP
3: Select MCC to negotiate
4: for all MCC selected do
5:    Ask MCC to start negotiation
6:    Create UCP(ResourceRequested)
7:    Ask UCP to start the negotiation
8:    if NegotiationAgreement = TRUE then
9:       Collect ResourceDataset from UCP
10:      Send ResourceDataset to the User or UCP
11:   end if
12: end for
```

Algorithm 2 Multicast Contributor Agent *(MCC)*

Inputs

ResourceOffered from the *User*
Constraint from the *User*
Request from the MCP
NegotiationAgreement from the UCC
ConstraintDataset from the UCC

```
1: Add MCC to the YP
2: while User does not stop the MCC do
3:    Get Request from some MCP
4:    if Request = ResourceOffered then
5:       if Child-Loop detected then
6:          Create UCC(Request,NUL)
7:       else
8:          Create UCC(Request,Constraint)
9:       end if
10:      Ask the UCC to start the negotiation
11:      if NegotiationAgreement = TRUE then
12:         Collect ConstraintDataset from UCC
13:      end if
14:   end if
15:   Remove UCC
16: end while
17: Remove MCC from the YP
```

Algorithm 3 Unicast Contributor Agent *(UCC)*

Inputs

Request from UCP
Constraint from MCC
ResourceOffered from MCC
ConstraintDataset from UCP
ConstraintSolved from UCP

```
1: if Constraint = NUL then
2:    Send ResourceOffered to UCP
3:    NegotiationAgreement ← TRUE
4: else
5:    Ask UCP to solve the constraint
6:    if ConstraintSolved = TRUE then
7:       Collect ConstraintDataset from UCP
8:       Send ConstraintDataset to MCC
9:       Send ResourceOffered to UCP
10:      NegotiationAgreement ← TRUE
11:   else
12:      NegotiationAgreement ← FALSE
13:   end if
14: end if
15: return NegotiationAgreement
```

5.3.1 Agreement Paths

After the activation of a new request by the user a *Request* object is created. An instance of this object will be linked to every UCP and includes i) the ID of the requesting node, ii) the ID of the first MCP Agent of the negotiation chain, and iii) the ID of the negotiating branch. The value of the ID of the negotiating branch corresponds to a list of numbers that increases at every step of the path creation. When an MCP is launched by a UCP it receives from its creator the *Request* object and adds to the branch ID a new number. In doing so the *Request* object will contain the information needed to create the agreement paths.

A UCP arrives to the end of a path candidate to solve a multilateral agreement, when it receives the requested data from its UCC without the need to launch any other MCP. Consequently, it creates a message that will represent the negotiation path to which the UCP belongs to. This object is propagated to the higher levels of the Petitioners chain up to the MCP that initiated the request. During this bottom up process of transferring the agreement path candidate, all the Petitioners, at every link of the path, add to the object the relevant information and reference of the nodes to which there is a possible agreement. These correspond to the nodes where the MCC participating in the negotiation process with every MCP are hosted.

At the end of the process of network exploration, the MCP that initiated the request receives, for every dataset of interest, the set of negotiation paths that correspond to a possible multilateral agreement. At that point, the MCP decides which negotiation paths to select from the possible candidates. An initial selection is performed among the paths that arrive to the same dataset, but the MCP may also decide to execute only a subset of all the remaining negotiation path candidates, based on other criteria (e.g. cost or reputation).

Algorithm 4 Unicast Petitioner Agent *(UCP)*

Inputs

ResourceRequested from MCP
constraint from UCC
constraintDataset from MCC

```
1: Ask UCC to send the ResourceRequested
2: if Constraint ≠ NUL then
3:    Search MCC in the Node to solve the constraint
4:    if  MCC ≠ NUL then
5:       ConstraintSolved ← TRUE
6:       Get ConstraintDataset from MCC
7:       Send ConstraintDataset to the UCC
8:    else
9:       Create MCP to look for the ConstraintDataset
10:      if  ConstraintDataset found then
11:         ConstraintSolved ← TRUE
12:         Send ConstraintDataset to the UCC
13:      else
14:         ConstraintSolved ← FALSE
15:         Notify failure to solve the constraint to the UCC
16:      end if
17:   end if
18: end if
19: if Constraint = NUL or ConstraintSolved then
20:    Collect ResourceRequested from UCC
21:    Send ResourceRequested to MCP
22:    NegotationAgreement ← TRUE
23: else
24:    NegotationAgreement ← FALSE
25: end if
26: return  NegotationAgreement
```

5.3.2 Loop Detection

Each agreement path or branch of the *Petitions Tree* is built during the network exploration. Every branch is identified with a *Request Identifier* corresponding to an array where each of its elements represent the participation of a Petitioner in the branch. It is important to note that an MCP will belong to more than one branch when i) it has more than one UCP exploring different options of agreement or ii) there is another MCP in the lower levels of its path with the same situation (managing more than one UCP). A new request received by an MCC is processed and compared with all the other active requests managed by the MCC.

A loop is identified when all the elements of the array of some *Request Identifier*, that are active in the MCC, is equal to the first elements of the *Request Identifier* of the new request received, which means that the request comes from the same branch of that already active request at the MCC. In that case, the associated UCC will be created without any constraint. Security issues that arise here have been studied and analysed [17].

Chapter 6

Improvements

"Strive for continuous improvement, instead of perfection"
Kim Collins

6.1 Key Performance Indicators

The main metric or *Key Performance Indicator (KPI)* to measure the level of success of the MOSAIC protocol execution is the total number of cases collected. Linked to this one, another *KPI* is the number of agreements achieved. Besides this, the efficiency of the protocol is also relevant and for this, we can measure the number of messages transmitted in the network and analyse the ratio of messages per agreement and case collected. All these metrics depend on the specific characteristics of the MOSAIC protocol, but also on the specific configuration of the network where MOSAIC operates.

6.1.1 Network properties

The performance of MOSAIC depends on the scenario where will be deployed. On the one hand, a collaborative network where nodes are offering significant amounts of data and achievable constraints, facilitates MOSAIC to find agreements for the multilateral exchange of the data. On the other hand, a network with nodes reluctant to share data, with difficult constraints to fulfill, makes the objective to build agreements more challenging.

These characteristics can be measured and formalised by the most relevant metrics of a network, that are the *degree* of its nodes, its *density*, and its *inclusiveness*.

Let's define the following key elements of a MOSAIC's network:
C and P: Total number of MCC (C) and MCP (P) activated in the network
$e_{i,j}$: The edge connecting MCC_i and MCP_j (MCC_i offers what MCP_j requests)
EP_i: Set of MCP_j connected to MCC_i (there is an edge $e_{i,j}$ connecting MCP_j and MCC_i)

EC_i: Set of MCC_j connected to MCP_i (there is an edge $e_{i,j}$ connecting MCC_j and MCP_i)
EC: Set of MCC_i where EP_i is not empty,
EP: Set of MCP_i where EC_i is not empty

Degree

The *degree of a node* (*deg*) is defined by the number of edges connected to it. In MOSAIC we can distinguish between the MCC *degree* (number of MCP requesting what a MCC offers), and the MCP *degree* (number of MCC offering what a MCP requests). The MCP *degree* is a metric that can be used to measure the difficulty to collect certain type of data. As higher MCP *degree*, as easier will be to collect the requested data. The MCC *degree* is a metric that can be used to measure the relevance of certain type of data. Higher values of MCC *degree* mean that the data offered is more demanded. Based on this, we can define C_{deg} as the average degree of MCC and P_{deg} as the average degree of MCP (see Eq. 6.1 and 6.2).

$$C_{deg} = \frac{\sum_{i=1}^{C} |\mathrm{EP}_i|}{C} \tag{6.1}$$

$$P_{deg} = \frac{\sum_{i=1}^{P} |\mathrm{EC}_i|}{P} \tag{6.2}$$

Inclusiveness

The *inclusiveness* of a network (*inc*) is a metric to measure the level of connected nodes, where *inc=1* refers to a network where all nodes have some connection, and *inc=0,5* means that a half of the nodes are isolated (with no connections to any other node). In MOSAIC's network we define two *inclusiveness* values (C_{inc} and P_{inc}), considering the number of connected MCC and MCP (see Eq. 6.3 and 6.4).

$$C_{inc} = \frac{|\mathrm{EC}|}{C} \tag{6.3}$$

$$P_{inc} = \frac{|\mathrm{EP}|}{P} \tag{6.4}$$

Density

The *density* of a network (*den*) is defined as the ratio of the existing edges and the theoretical total number of edges possible. In MOSAIC's network, a theoretical configuration where every MCC offers all data requested by all MCP, corresponds to the maximum number of connections possible: $C \times P$.

Table 6.1: Network type based on the behaviour of MCC and MCP

	Data offer high	**Data offer low**
Data requests high	**Network A** Highly collaborative and active network. All inclusiveness and density values are high.	**Network B** Poor collaborative network, but active in data search. *Pinc* and *Pden* are low, and *Cinc* and *Cden* are high.
Data requests low	**Network C** Highly collaborative network, but inactive in data search. *Pinc* and *Pden* are high, and *Cinc* and *Cden* are low.	**Network D** Poor collaborative and inactive network. Nodes reluctant to share data and little data access requests. All inclusiveness and density values are low.

Two *density* values can be defined: C_{den} to measure the density of the MCC connections (see Eq. 6.5) and P_{den} to measure the density of the MCP connections (see Eq. 6.6).

$$C_{den} = \frac{\sum_{i=1}^{|\mathrm{EC}|} |\mathrm{EP}_i|}{|\mathrm{EC}| \cdot P} \tag{6.5}$$

$$P_{den} = \frac{\sum_{i=1}^{|\mathrm{EP}|} |\mathrm{EC}_i|}{|\mathrm{EP}| \cdot C} \tag{6.6}$$

The global density in MOSAIC's network is represented in Eq. 6.7.

$$den = \frac{\sum_{i=1}^{|\mathrm{EC}|} |\mathrm{EP}_i|}{C \cdot P} \tag{6.7}$$

Characterising MOSAIC's network

According to the levels of data offers and demands, there are four categories of networks (see Table 6.1). As more collaborative the nodes of a network are (types A and C), easier to reach a multilateral agreement is, shorter the path (lower TTL values) required and less branches (MCC candidates) are required to explore.

As less collaborative the nodes of a network are (types B and D), less relevant is to build complex strategies for the selection of the path (there will be less MCC to choose from during the exploration) and more MCC can be selected for the path exploration.

From the perspective of the MOSAIC actors (MCP and MCC), the topology of the network is unknown and there is no preliminary knowledge about the type of network where the protocol operates. There is however the possibility to learn from the experience after a number of exploration requests and infer the class of the network where the protocol is working in, and parametrising it according to this.

6.1.2 Protocol properties and theoretical mathematical model

In addition to the general metrics to characterise a network, the MOSAIC protocol has a set of specific metrics that will allow a quantitative validation (see chapter 7), assess its performance and efficiency, and demonstrate the viability of the MOSAIC deployment in the real world. The main *KPI* of MOSAIC are i) the number of messages transmitted, ii) the number of agreements achieved and iii) the number of cases collected.

Messages

During the negotiation process the agents involved generate a number of messages. Those that correspond to the communications among the agents involved in a successful multilateral agreement i, correspond to the cardinality of MSG_i (see Eq. 6.8).

$$|\mathrm{MSG}_i| = |\mathrm{M}_i| \times (4\,|\mathrm{U}_i| + 2) \tag{6.8}$$

Where:
msg_i: Message used in the dialog performed to achieve the agreement i
MSG_i: Set of msg_i, belonging to the communication required to reach agreement i
M_i and U_i: Set of MCP and UCP participating in the agreement i

The two messages per MCP correspond to:
MCP $\rightarrow$ **YP**: Request from MCP to YP asking for MCC offering the desired dataset
YP $\rightarrow$ **MCP**: Response from YP to MCP with the list of references of MCC available

The four messages per UCP correspond to the following:
MCP $\rightarrow$ **MCC**: Dataset request
UCC $\rightarrow$ **UCP**: Notification of the constraint
UCP $\rightarrow$ **UCC**: Constraint delivery
UCC $\rightarrow$ **UCP**: Acceptance of the agreement

MOSAIC's efficiency can be measured by the number of messages required and transmitted among the MOSAIC's actors in order to reach a multilateral agreement. The difficulty to collect certain data requested by a MCP depends on a number of variables and the efficiency of the protocol differs also for each MCP. Let's define e_i, as the efficiency of the protocol for MCP_i, according to Eq. 6.9, where n is the number of agreements achieved by MCP_i.

$$e_i = \frac{1}{n} \sum_{j=1}^{n} |\mathrm{MSG}_j| \tag{6.9}$$

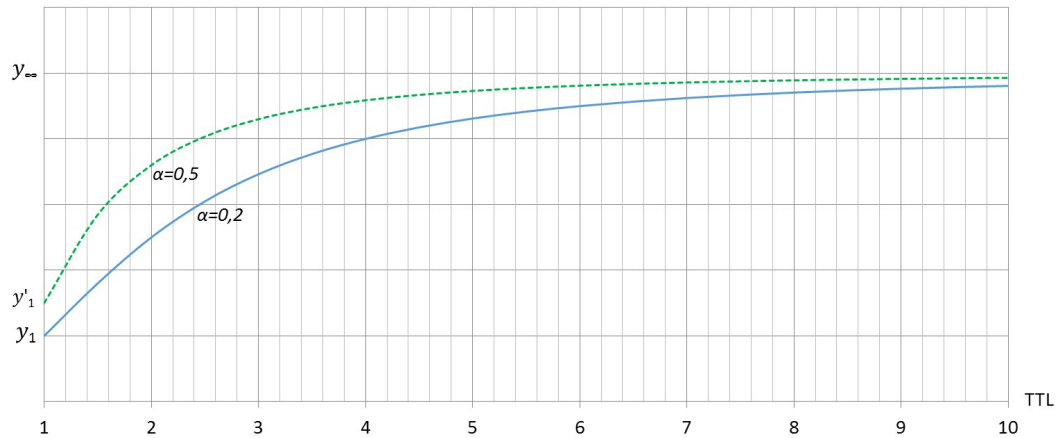

Figure 6.1: Representation of Eq. 6.10. The total number of agreements in the network is expected to increase as larger is the path allowed to build a collaboration. y_1 is the total number of bilateral agreements (using TTL=1) and y_∞ is the maximum number of multilateral agreements possible. Different strategies for the network exploration are represented (one with a solid line and a more effective one with a dotted line), providing different shapes of the curve.

Multilateral agreements

It is expected that the number of agreements achieved increases as larger is the number of nodes allowed to participate in the multilateral exchange. This is represented by $\hat{y}$ (see Eq. 6.10), an equation that models the behaviour of MOSAIC, calculating the theoretical values of one of its main *KPI*. The shape of this function is shown in Figure 6.1.

$$\hat{y} = y_\infty - \frac{y_\infty - y_1}{1 + \alpha\,(x^2 - 1)} \tag{6.10}$$

Where:
$x \in \{1, \ldots, \mathrm{N}\}$: Maximum length of the path allowed. It is equivalent to the TTL parameter. N corresponds to TTL_{max} (the maximum length of a path). N+1 is the maximum amount of nodes participating in an agreement
y_1: Number of agreements achieved through bilateral collaborations (TTL=1)
y_∞: Maximum number of agreements possible to achieve through multilateral collaborations
α: Coefficient to adjust the grow speed of $\hat{y}$. ($\alpha > 0$)

Cases collected

The main *KPI* to measure MOSAIC's performance is the number of cases collected from the network which directly depends on the number of agreements achieved.

The number of cases collected starts with those coming from bilateral agreements and will increase as larger length of the multilateral agreement is allowed, approaching to the limit corresponding to the total number of cases available in the network (of the requested type).

It is expected that the amount of cases collected will grow rapidly during the the initial extensions in the number of nodes allowed to participate in the multilateral agreements and when most of the cases available are already collected, the increase in the number of nodes allowed to participate in the agreements will have less and less impact. The shape of this grow is expected to be similar to Fig. 6.1 following also Eq. 6.10 (replacing *number of agreements* by *number of cases collected* and adjusting the value of coefficient α).

The performance of different strategies for the network exploration can be measured according to the number case collected per agreement.

An update of the efficiency of the protocol calculated in Eq. 6.9 (according to the messages transmitted per agreement) is to consider the efficiency, not per agreements achieved, but per case collected. According to this, let's define e'_i as the efficiency of the protocol for MCP_i, according to Eq. 6.11, where n is again the number of agreements achieved by MCP_i and D_j is the data collected (in number of cases) per each agreement.

$$e'_i = \frac{\sum_{j=1}^{n} |\mathrm{MSG}_j|}{\sum_{j=1}^{n} \mathrm{D}_j} \tag{6.11}$$

Complexity

The complexity (and also the performance) of MOSAIC depends on the number of branches allowed to explore by every MCP (w), that corresponds to the number of MCC selected among the candidates offering what MCP requests, and the maximum length of the paths (TTL). Let's RP_r be the set of MCP activated for a single request r. Its maximum is calculated in Eq. 6.12.

$$\max(|\mathrm{RP}_r|) = \sum_{n=1}^{\mathrm{TTL}} w^n \tag{6.12}$$

The computational cost of MOSAIC is exponential and in the worst case corresponds to $O((w|\mathrm{EP}|)^{\mathrm{TTL}})$. Although actual number of active MCP and MCC in the network reduces this theoretic maximum cost, there is a crucial need for a proper selection of the path during the network exploration.

This cost could be compared with the complexity of the Dijkstra's algorithm to find the shortest path in a network, which is $O(E + V\log(V))$, where V is the number of *vertex* and E the number of *edges*. In our case, $V = |\mathrm{EC}| + |\mathrm{EP}|$ and and $\max(E) = |\mathrm{EC}| \times |\mathrm{EP}|$. It is clear that Dijkstra's algorithm has a better performance, but can be only used when the topology of the network is known, which is not the case of the scenarios where MOSAIC aims to operate.

6.2 Strategies for the network exploration

6.2.1 Criteria for the path selection

The goals for an intelligent selection of the path for a multilateral agreement are: i) get as much data as possible from the network, ii) get the most appreciated data (ranked with higher quality marks) and iii) reduce the risk of agreement failure (rollback). To this end, the selection of the best MCC by the MCP among all candidates is crucial. Let's describe some of the criteria that can be used for the path selection during the exploration of the network.

Random selection

The basic initial criteria for the path selection is the random selection of one MCC among all candidates. More advanced and complex criteria can be compared against this one as the reference and baseline criteria in order to analyse and measure the improvements provided by the others.

Dataset size

A Node with a big dataset (including a significant number of cases of different data type), first has likely more cases of the requested data and perhaps of higher quality, which is of course of the interest of the petitioner node. And second, when the agreement path exploration finds a loop, the node with a large dataset has more chances to has the requested data of the last MCP, solving automatically the constraint and making possible the agreement (see section 5.1 "Stage 2: Network exploration"). For these two reasons, the *Dataset size* of the node can be a criteria for the selection of a MCC among all candidates.

Agreement Reputation

The *Agreement Reputation* of a Node can be obtained through the right behaviour of the MCC that should fully respect their commitments when establishing an agreement with a MCP. During the process of data exchange initiated after the achievement of an agreement, each UCP involved notifies its MCP about the fulfilment of the commitments from the associated UCC. At the end, this concludes with the commit (successful transaction) or rollback (failed transaction) of the whole multilateral agreement. If the data is correctly transmitted and corresponds to the type of data published, then the Node whose UCP sent the right data will increase its reputation. The Nodes of MCC with bad behaviour (not transferring the data committed, or transferring data of another data type, or with low quality) can see reduced their reputation or even banned for further occasions.

All Nodes involved in an agreement know about the behaviour of the rest of the Nodes and can keep record of their marks. The level of the reputation of all Nodes can be notified to the MCP when is activated, and considered to select a MCC or another.

Dataset reputation

After a successful data exchange agreement (and transaction), every MCP provides a rank of the data delivered by the MCC. Each score is kept at the Node level of the MCP and can be shared with the other participating Nodes of the agreement. The score for the data collected can be a combination of size, quality, singularity and cost (if any).

Each Node has different sizes, qualities and costs per data type. Thus, there may be a not a single dataset reputations per Node, but a score per each data type and Node. Similarly to the agreement reputation, the *Dataset Reputation* of the Nodes can be notified to the MCP when is activated, and the score for the same dataset type that the MCP looking for can be considered to select a MCC or another.

Trust

Another criteria for the selection of a MCC is *Trust*. *Agreement* and *Dataset Reputation* is a component of trust, but not the only one. Taking advantage from previous research in agent trust, argumentation and reasoning [99, 87, 27], we can combine and include the opinions from other Nodes when calculating the trust in a MCC. The relevance of other's opinions towards a Node (or its MCC), can be adjusted to different parameters, including the similarities we have with them and our own trust towards them.

When a new Node joins the network, its reputation is still low. Nevertheless, we can design the MOSAIC protocol to assume some risk, give some trust to unknown nodes, and choose their MCC if there is no clear alternative choice.

6.2.2 Combined intelligent analysis

As each criteria described above can have more or less relevance a weight (p) can be assigned to each and calculate the global ranking (r_i) of a MCC$_i$ according to Eq. 6.13, order all MCC according to their associated r, and select the one with higher score.

$$r_i = \frac{1}{n} \sum_{j=1}^{n} p_j c_{ji} \tag{6.13}$$

Where:
r_i: Ranking of MCC$_i$
n: Number of criteria (Dataset size, reputation, ...)

$c_{ji} \in [0,1]$: Value of criteria j of MCC_i
$p_j \in [0,1]$: Weight of criteria j where $\sum_{j=1}^{n} p_j = 1$

Eq. 6.13 proposes an integrated use of all criteria to select the MCC and build the path for a multilateral agreement. However, we have seen that some criteria have some correlation with some other (e.g. reputation and trust) and it is not clear how to set the precise weight to each of them. Finding the relevant criteria and the best combination in order to select the best path can be performed using machine learning techniques and case based reasoning.

Data preparation for the Machine Learning analysis

The Machine Learning process needs data marked with every class. In our case:

Class 1: Path explorations that conclude with a successful multilateral agreement
Class 2: Path explorations that conclude with a failed attempt of multilateral agreement.

A scenario where a number of explorations are prepared for execution has to be defined. Then, the MOSAIC protocol has to be executed and the data generated, stored for its further analysis. During the protocol's execution, each record of the data to store for further analysis has to include the reference of the MCC selected, plus all the data linked to the exploration, specially the reference of the criteria used for the selection of the MCC (as described above), but also other metrics that may become relevant. When the exploration concludes, each of these records has to be marked with the corresponding class it belongs to (1: successful exploration or 2: failed exploration).

Creating a classifier

Different methods of Machine Learning can be used, namely those to build decision trees like J48 and Random Forest, methods to generate probabilistic classifiers like Naïve Bayes, and methods like Support Vector Machines that distinguish between existing classes and allocate new cases into the class that better fits. A significant part of the data prepared in the first step (about 90% of records) will feed the analysis process. This process will identify relevant metrics to be used by the classifiers which will predict the class that a record belongs to.

Selecting the best classifier

The portion of the data set not used for training the classifiers (about 10% of the records) is used for their validation and to calculate their accuracy. Per each of those records each classifier will infer whether the exploration will conclude with a successful multilateral agreement or on contrary will be a failed attempt. The result of the classifier will be compared with the actual

and real mark that shows whether the exploration attempt really belongs to Class 1 or to Class 2, and with this the accuracy, sensitivity (% of true positives) and specificity (% of true negatives) of the classifier can be calculated.

- **Sensitivity:** A *true positive* means that the classifier predicted that the selection of a certain MCC will conclude with a successful agreement and this actually was the result of this path selection. A *false positive* occurs when the classifier predicted that the selection of a certain MCC will conclude with a successful agreement, but in real, the selection of that MCC concluded with a failed attempt to build a multilateral agreement. The percentage of *true positive* corresponds to the *sensitivity* value of a classifier.

- **Specificity:** A *true negative* means that the classifier predicted that the selection of a certain MCC will conclude with a failed attempt of a multilateral agreement, and this actually was the result of this path selection. A *false negative* occurs when the classifier predicted that the selection of a certain MCC will conclude with a failed attempt of a multilateral agreement, but in real, the selection of that MCC concluded with a successful multilateral agreement. The percentage of *true negatives* corresponds to the *specificity* value of a classifier.

- **Accuracy:** The percentage of right predictions of the classifier (either true positives or true negatives) is the accuracy.

In MOSAIC, more than the overall *accuracy*, the most relevant value to select the best classifier is the *sensitivity* as we aim to maximise successful agreements.

Using the classifier

Once the classifier is ready, all data prepared for its development can be removed. The classifier is a formula and its variables data that can be processed in real time for a MCP at the moment of selecting the MCC.

6.2.3 Dynamic, real time and distributed analysis

A real time update of the classifiers is possible using *Stream Data Mining* techniques. This permanent update and self-learning of the classifiers according to the results of their predictions, permits to adapt their behaviour to changing networks and keep their performance at the maximum level possible.

Due to the distributed nature of MOSAIC it is not possible to build global classifiers to operate in the whole network. Traditional Machine Learning techniques or more sophisticated Stream Data Mining algorithms have to be executed at Node level. Each classifier can be created

from the perspective and experience of each Node and Classifiers themselves may become an asset that could be part of an exchange agreement.

6.2.4 MOSAIC's strategy for the first deployment

The MOSAIC protocol is designed to support a number of techniques and strategies for the best selection of a path in order to maximise the chances to achieve a multilateral agreement. The algorithm that implements the strategy to select one MCC or another can be added into the protocol producing a number of versions of it.

For the first deployment and evaluation of the protocol, the criterion of the path selection chosen is the size of the dataset hosted at the MCC node. This criterion will be compared with the random selection of an MCC among the list of candidates and the results are described in the evaluation of the protocol in the following section 7.

Chapter 7

Validation and Evaluation of MOSAIC

"Before software can be reusable it first has to be usable"
Ralph Johnson

The assessment performed to the MOSAIC system includes the validation and evaluation of i) the correctness of the protocol, ii) the advantatges of multilateral agreements compared with bilateral ones; iii) the optimisation process for the network exploration; and iv) the analysis of the type of nodes that most benefit from the MOSAIC system. For this evaluation two criteria for the network exploration are compared: i) A random selection of the path and ii) A selection of the path based on the dataset size of the node.

The figures obtained from the empirical execution of the protocol will be also compared with the theoretical values generated by the mathematical model formalised in section 6 (Eq. 6.10). Doing so, specific values of the parameters of the model will be calculated, being them precise mesures of the efficiency of the protocol and useful for comparing different strategies.

7.1 The Scenario evaluation

The database of the evaluation scenario

The scenario evaluation corresponds to a set of nodes (cities) hosting each of them a number of datasets with clinical cases of brain tumours (see Table 7.1). While some datasets are freely offered to the network without any restriction, most of them have a constrain associated, requiring the delivery of some other dataset from some other node.

For a better prognosis of a certain patient, analysing the information collected from other patients with similar profiles and suffering the same tumour type may be relevant. Moreover, the knowledge of the effect of certain therapies to other patients may help the clinician to provide a more effective and personalised treatment. The MOSAIC system could help both prognosis and theragnosis by facilitating the multilateral agreements for data exchange.

For each dataset each node activates an MCC. Every MCC is associated to a constraint

Table 7.1: Dataset of the scenario evaluation. Brain tumours by major histology groupings

Brain and CNS tumours			**Worldwide 2.852 Nodes**		**USA 205 Nodes**	
Class and Histology		**Types**	**Datasets**	**Cases**	**Datasets**	**Cases**
A	tumours of Neuroepithelial Tissue	A1-A17	8.203	374.580	859	41.360
B	tumours of Cranial and Spinal Nerves	B1	2.145	114.680	201	12.750
C	tumours of Meninges	C1-C3	3.135	507.800	238	54.380
D	Lymphomas and Hematopoietic Neoplasms	D1	759	23.610	95	2.690
E	Germ Cell tumours and Cysts	E1	124	2.310	14	240
F	tumours of Sellar Region	F1-F2	2.947	188.060	227	20.360
G	Local Extensions from Regional tumours	G1	5	50	1	10
H	Unclassified tumours	H1-H2	1.584	58.980	189	6.720

corresponding to a brain tumour randomly selected from all possible types - with the same probability as any data type - or to an empty constraint, in which case the MCC freely offers its dataset to any MCP. Two datasets have been created: One with 2.852 nodes corresponding to the main cities around the world, hosting 18.902 data sets; and another one for the most complex and time consuming evaluations with a subset of 205 cities with 1.824 datasets.

The evaluation of MOSAIC has been performed on this simulated and realistic scenario. The results shown are based, first on the activation of the MCC for the datasets with the same cases available, and second on the activation of a request (or MCP) for every possible dataset, by every node (or city). This corresponds to $2.852 \times 28 = 79.856$ requests for the whole worldwide network and to $205 \times 28 = 5.740$ requests for the smaller network.

The CBTRUS Statistical report provides the figures of the total number of Brain and CNS tumours in USA by major histology groupings (see Table 7.2) and a simulated number of cases per node has been calculated according to the population of every city and the resulting proportional figure considering the USA population and the number of tumour cases from the CBTRUS DB. Table 7.3 shows a subset of this DB for the first 25 cities and 17 tumour types (from the total number of classes).

The constraints have been simulated calculating a random figure (from 0 to 29) at every node for every dataset. '0' represents that there is no dataset available of that tumour type and no constraint can be assigned for delivering nothing. Any number between '1' and '28' indicates the reference of the tumour type to be delivered by the requesting node (as constraint for authorising the access to the data). '29' indicates that there is no constrain to fulfill and the cases available at the node for that specific tumour type will be freely delivered to the requesting node. Table 7.4 shows a subset of these constrains for the first 25 cities and 17 tumour types[1].

[1]The complete DB used for the simulations from which the figures of tables 7.3 and 7.4 have been obtained is freely available upon request or can be also downloaded from the author's site at www.researchgate.net

Table 7.2: Total amount of cases of every data type

Type	Histology	Cases
A	**tumours of Neuroepithelial Tissue**	**76.340**
A1	Pilocytic astrocytoma	3.663
A2	Protoplasmic and fibrillary astrocytoma	1.242
A3	Anaplastic astrocytoma	4.747
A4	Unique astrocytoma variants	1.097
A5	Astrocytoma, NOS	5.194
A6	Glioblastoma	37.890
A7	Oligodendroglioma	3.184
A8	Anaplastic oligodendroglioma	1.386
A9	Ependymoma/anaplastic ependymoma	3.011
A10	Ependymoma variants	1.135
A11	Mixed glioma	2.251
A12	Glioma malignant, NOS	4.963
A13	Choroid plexus	511
A14	Neuroepithelial	236
A15	Non-malignant and malignant neuronal/glial	3.169
A16	Pineal parenchymal	404
A17	Embryonal/primitive/medulloblastoma	2.257
B	**Cranial and Spinal Nerves**	**19.605**
B1	Nerve sheath	19.600
C	**tumours of Meninges**	**80.457**
C1	Meningioma	77.908
C2	Other mesenchymal	667
C3	Hemangioblastoma	1.882
D	**Lymphomas and Hematopoietic Neoplasms**	**5.380**
D1	Lymphoma	5.380
E	**Germ Cell tumours and Cysts**	**1.092**
E1	Germ cell tumours, cysts and heterotopias	1.092
F	**tumours of Sellar Region**	**31.405**
F1	Pituitary	29.806
F2	Craniopharyngioma	1.599
G	**Local Extensions from Regional tumours**	**194**
G1	Chordoma/chondrosarcoma	194
H	**Unclassified tumours**	**12.318**
H1	Hemangioma	1.814
H2	Neoplasm, unspecified	10.373
H3	All other	131
	TOTAL	**226.791**

Table 7.3: Cases per node (city) and tumour type from the DB used for the simulations (sample). For example: At the Atlanta node (row number 11), there are 30 cases of tumour type A1 and 10 cases of tumour type A2

City		**Brain tumour type**																
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17
1	Akron	0	0	10	0	10	90	0	0	0	0	0	10	0	0	0	0	0
2	Albany	10	0	10	0	10	120	10	0	0	0	0	10	0	0	10	0	0
3	Albuquerque	0	0	10	0	10	90	0	0	0	0	0	10	0	0	0	0	0
4	Allentown	0	0	10	0	10	80	0	0	0	0	0	10	0	0	0	0	0
5	Amarillo	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0
6	Anaheim	30	10	40	10	50	390	30	10	30	10	20	50	0	0	30	0	20
7	Anchorage	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0
8	Ann Arbor	0	0	10	0	10	70	0	0	0	0	0	10	0	0	0	0	0
9	Appleton	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0
10	Asheville	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0
11	Atlanta	30	10	40	10	50	370	30	10	20	10	20	40	0	0	30	0	20
12	Atlantic City	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0
13	Augusta	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0
14	Austin	10	0	20	0	20	170	10	0	10	0	10	20	0	0	10	0	10
15	Bakersfield	0	0	10	0	10	90	0	0	0	0	0	10	0	0	0	0	0
...																		

Table 7.4: Constrains per node and dataset (sample). The number in each cell corresponds to the reference of the dataset that needs to be given (from A1 to H3) in order to allow access to the corresponding dataset. For example: At the Atlanta node (row number 11), the constraint to give access to dataset A1 (Pilocytic astrocytoma) is the prior delivery of a dataset type F2 (Craniopharyngioma), with reference number 25

City		**Brain tumour type**																
		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17
1	Akron	0	0	4	0	26	5	0	0	0	0	0	5	0	0	0	0	0
2	Albany	3	0	16	0	2	4	7	0	0	0	0	18	0	0	5	0	0
3	Albuquerque	0	0	25	0	1	25	0	0	0	0	0	2	0	0	0	0	0
4	Allentown	0	0	6	0	22	5	0	0	0	0	0	11	0	0	0	0	0
5	Amarillo	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	Anaheim	21	3	5	27	6	3	10	19	18	2	5	25	0	0	25	0	4
7	Anchorage	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0
8	Ann Arbor	0	0	27	0	11	9	0	0	0	0	0	10	0	0	0	0	0
9	Appleton	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0
10	Asheville	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0
11	Atlanta	25	12	1	14	5	28	11	1	13	15	9	13	0	0	16	0	7
12	Atlantic City	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0
13	Augusta	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0
14	Austin	16	0	17	0	26	16	18	0	6	0	4	6	0	0	13	0	15
15	Bakersfield	0	0	2	0	9	26	0	0	0	0	0	23	0	0	0	0	0
...																		

Table 7.5: Data Base generated after the protocol execution, where a Node requests cases of a certain data type. 'C' corresponds to the initial number of cases of the type requested hosted at the requesting node, 'MCP' is the number of Multi-cast Petitioner Agents participating in the multilateral agreement, 'MSG' the number of messages exchanged and 'Path' the average length of the multilateral agreement paths achieved.

Requesting Node	**Requested Data**	**C**	**MCP**	**MSG**	**Path**	**Cases collected**
Akron	D01	0	2	20	3	10
Akron	D02	0	6	156	7	10
Akron	D03	10	2	20	3	70
Akron	D04	0	8	272	9	10
Akron	D05	10	7	210	8	10
Akron	D06	90	6	156	7	40
...	...	...	...	...	...	...

On one hand, nodes with a large number of cases covering most of the data types will have a higher chance to directly solve a possible constraint and achieve bilateral agreements. On the other hand, the nodes with a reduced number of cases in their datasets will likely need multilateral agreements to get the data desired from the network. One of our hypothesis is that MOSAIC will be especially useful for nodes with less chances to achieve bilateral agreements.

Simulation output

After the simulation execution a DB with the total number of cases collected per node and datatype was created. Table 7.5 shows a subset of that database after the execution of the simulator with TTL=20. All the results presented in this section have been obtained after the processing of these figures.

7.2 Correctness of the MOSAIC protocol

A preliminary evaluation of the MOSAIC protocol is a cross validation to check that it works properly. For this, algorithm 5 has been created. It scans the network of 2.852 cities and their 18.902 data sets seeking for all possible bilateral agreements. This algorithm generates as an output a matrix with the figures corresponding to all the cases collected by each node for every data type after the bilateral data exchanges. These figures have been compared with those obtained by the execution of the MOSAIC protocol with the Time to Live parameter set to 1, forcing that the maximum length of every multilateral agreement is limited to 2 nodes. The results obtained in both cases are exactly the same (see Figure 7.1), showing the correctness of the protocol for this scenario.

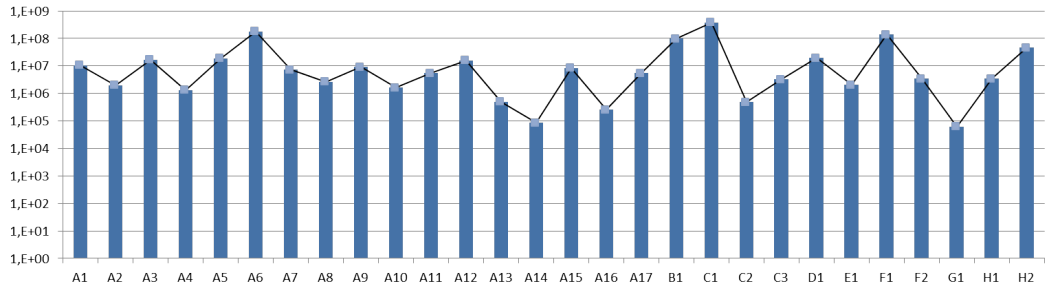

Figure 7.1: Cases collected after bilateral agreements, calculated to validate the correctness of MOSAIC using two methods: i) Algorithm 5 and ii) MOSAIC with TTL=1.

Algorithm 5 Search for bilateral exchange agreements

```
1: for i = 1 to numCities do
2:   for j = 1 to numDataTypes do
3:     for k = 1 to numCities do
4:       if i ≠ k then
5:         if Dataset[k, j] ≠ 0 then
6:           if constraint[k, j] = nul then
7:             collect Dataset[k, j] for node i
8:           else if constraint[k, j] available in node i then
9:             solve constraint from data in node i
10:            collect Dataset[k, j] for node i
11:          end if
12:        end if
13:      end if
14:    end for
15:  end for
16: end for
```

7.3 The impact of multilateral agreements

7.3.1 Improvements in number of agreements and cases collected

The first evaluation is to prove the main goal of the MOSAIC system which is to overcome the amount of data that can be exchanged with bilateral agreements and collect as much data as possible from the network by achieving as much data exchange agreements as possible.

Due to the time constraints during the simulations of the protocol behaviour with different TTL values, a subset of the whole DB has been created. Only 205 cities of the initial DB have been used for these simulations.

The results obtained strongly depend on the parameters of the protocol, namely its TTL

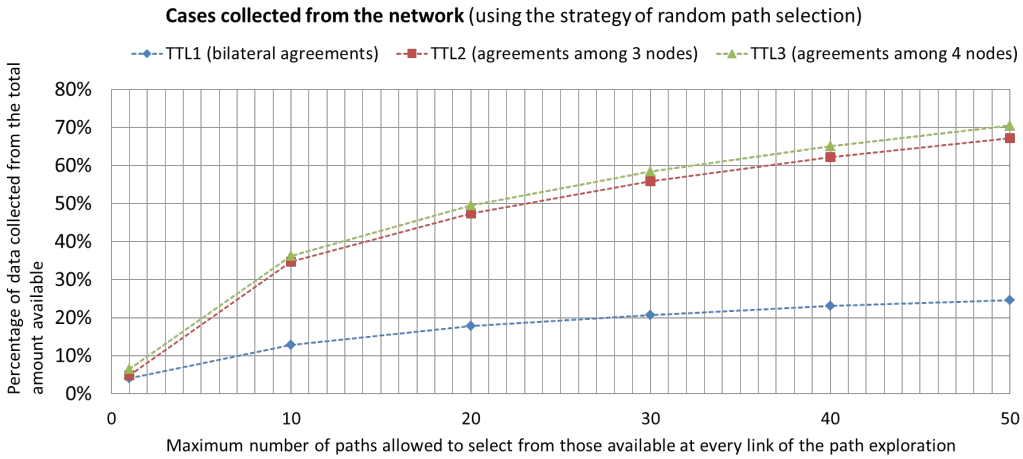

Figure 7.2: Percentage of cases collected from the total number available in the network with different values of TTL and size of the selected path set.

and the number of branches selected from all the paths available during the network exploration. Figure 7.2 shows the results with different values for TTL (1: bilateral agreements, 2: agreements among 3 nodes, and 3: agreements among 4 nodes) and with a range selection of paths starting from 1 (only exploring a single MCC from all available) to 50 (higher values of TTL are not needed as most of the data available in the network is made accessible after much more short negotiation paths). The percentages of data collected shows a steady increase when the selection of the number of possible paths increases, and while the improvement from TTL1 to TTL2 is significant, the increase from TTL2 to TTL3 is limited.

7.3.2 Evaluation of the path selection strategy and model's validation

The second evaluation refers to the optimisation process for the network exploration through the intelligent selection of the paths to follow. As indicated in section 5, the MCP receives the list of MCC compatible from the Yellow Pages and in order to avoid unmanageable network explorations the MCP has to decide which to select and which to discard. Two strategies have been evaluated. One selects a MCC randomly from those available, and the other selects the MCC with the biggest dataset. The two cases have been tested using the database of 205 cities with 1.824 datasets in total.

The number of agreements grows as higher are the values of the TTL parameter (see Figure 7.3). This improvement is also higher when selecting the path to follow during the network exploration according to the size of the MCC dataset (red dots), instead of a random selection among the MCC available (blue dots).

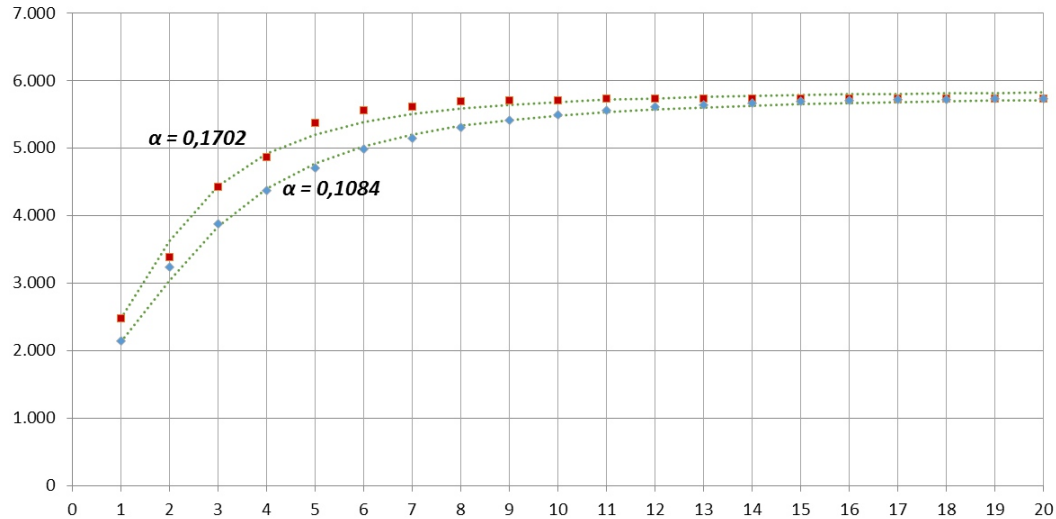

Figure 7.3: Total number of agreements using MOSAIC with different TTL values (from 1 to 20) and two strategies for the path selection: i) Random selection (blue dots) and ii) Selection of the node's path with largest data set size (red dots). Dotted lines come from the model formalised in section 6 (Eq. 6.10).

The results obtained for the evaluation of the MOSAIC strategy can be compared with the estimated figures generated by the Equation 6.10. The α parameter of this equation allows to adjust the shape of the curve to fit with the real figures obtained. For the case with a random selection of the path, $\alpha = 0,1084$ gives the closest estimation and for the case of the path selection based on the dataset size, $\alpha = 0,1702$ gives the best approximation to the real figures obtained. The value of α can be used as a metric for the measurement and compared analysis of the efficiency of different strategies of the protocol for the path's selection.

The average error of the predicted values for the number of agreements from the model (Eq. 6.10) and the actual figures obtained from the MOSAIC execution is 0,8% with a standard deviation $\sigma = 1,26\%$ for the random selection case and for the case of the path selection based on the dataset size, the average error is 1,4% with a standard deviation $\sigma = 1,54\%$.

The MOSAIC efficiency can be also measured with the number of messages required to achieve an agreement. Figure 7.4 shows the reduction in the number of messages transmitted over the network needed to achieve an agreement, comparing the selection criteria of the MCC between the strategy based on the dataset size and the random selection. While this figure is similar for low values of TTL, the difference increases when the length of the chain allowed per agreement grows. This analysis demonstrates that better criteria for the path selection improves not only the amount of data and agreements achieved, but also the efficiency of the protocol.

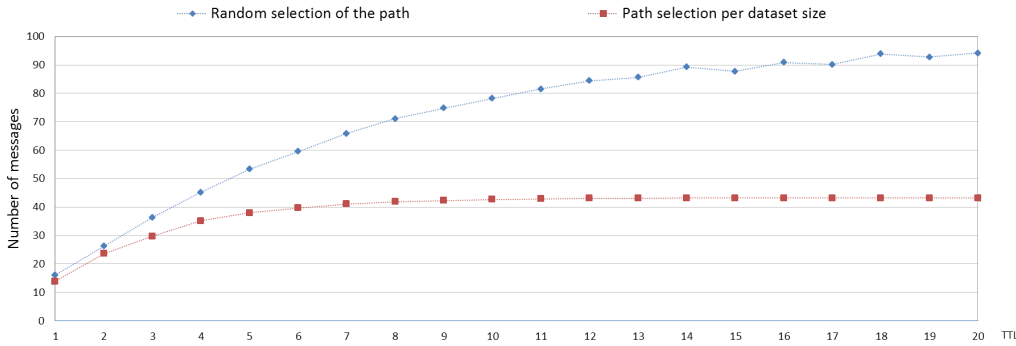

Figure 7.4: Comparative of the average number of messages needed to achieve an agreement, between the two branch selection strategies.

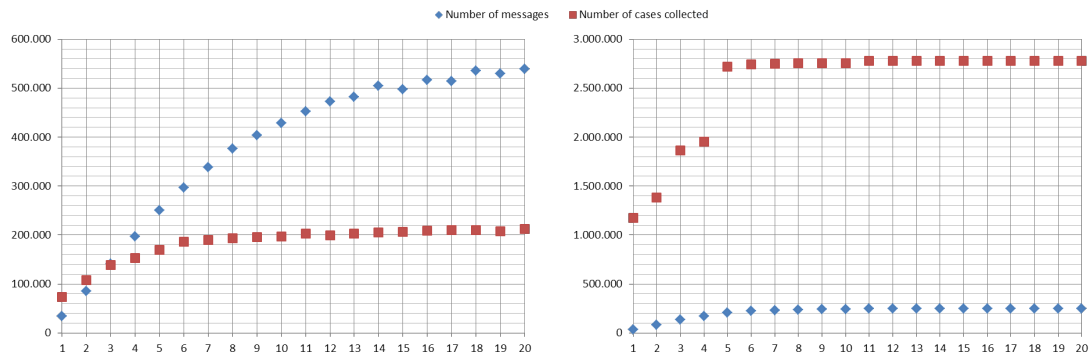

Figure 7.5: The total number of cases collected and messages transmitted with the two path selection strategies: Random selection (left) and selection based on the dataset size (right).

There are significant differences in the efficiency of the protocol both in terms of cases collected and messages transmitted when comparing the two strategies for the path selection. This can be seen in Figure 7.5 and Table 7.6. While the strategy of random path selection (left side of the figure) has a limit in the number of cases collected near to 0,2M, the strategy of the path selection based on the dataset size (right side of the figure) arrives to 2,8M cases (total number of cases collected by all MCP in the network). Moreover, when looking at the number of messages transmitted, although for lower values of the TTL parameter are similar between the two strategies of path selection, when the length of the path for the multilateral agreements grows, it is more or less twice the amount of messages transmitted in the strategy of the random selection.

Table 7.6: Values of the main metrics obtained from the MOSAIC execution showing the results based on the two selected strategies for path selection: random and dataset size. #MSG: Number of messages transmitted; #cases: Number of cases collected; MN: Maximum number of nodes involved in an agreement; Agr: Number of agreements; and C/M: Number of cases collected per message.

	RANDOM SELECTION					SELECTION PER DATASET SIZE				
TTL	**#MSG**	**#cases**	**MN**	**Agr.**	**C/M**	**#MSG**	**#cases**	**MN**	**Agr.**	**C/M**
1	34.436	73.324	2	2.133	2,13	34.436	1.173.240	2	2.475	34,07
2	84.735	107.932	3	3.225	1,27	80.132	1.381.080	3	3.389	17,24
3	140.894	138.950	4	3.869	0,99	131.832	1.863.890	4	4.427	14,14
4	197.025	152.912	5	4.360	0,78	171.192	1.953.260	5	4.864	11,41
5	250.507	169.580	6	4.698	0,68	204.442	2.720.580	6	5.380	13,31
6	296.234	185.860	7	4.971	0,63	220.956	2.745.110	7	5.563	12,42
7	338.551	190.110	8	5.141	0,56	230.460	2.748.810	8	5.610	11,93
8	375.922	193.530	9	5.289	0,51	238.458	2.755.690	9	5.696	11,56
9	404.063	195.862	10	5.404	0,48	241.468	2.756.690	10	5.708	11,42
10	429.165	196.616	11	5.484	0,46	243.886	2.756.730	11	5.712	11,30
11	452.606	203.126	12	5.549	0,45	246.208	2.778.490	12	5.732	11,29
12	472.320	199.396	13	5.596	0,42	246.866	2.778.490	12	5.732	11,26
13	482.090	202.732	14	5.631	0,42	247.580	2.778.930	14	5.738	11,22
14	504.650	205.352	15	5.653	0,41	247.690	2.778.930	14	5.738	11,22
15	497.940	206.458	16	5.677	0,41	247.808	2.778.940	16	5.739	11,21
16	517.087	209.304	17	5.691	0,40	247.808	2.778.940	16	5.739	11,21
17	514.562	210.036	18	5.707	0,41	247.808	2.778.940	16	5.739	11,21
18	536.030	209.370	19	5.710	0,39	247.808	2.778.940	16	5.739	11,21
19	530.238	207.642	20	5.716	0,39	247.808	2.778.940	16	5.739	11,21
20	538.571	212.780	21	5.721	0,40	247.808	2.778.940	16	5.739	11,21

7.3.3 MOSAIC performance per type of node

Finally, we validated the hypothesis that the nodes with less data (which have less chances to achieve bilateral agreements) would be those that specially benefit from MOSAIC. The figures that validate this hypothesis are shown in Figure 7.6, generated after running the MOSAIC protocol with TTL=1 (bilateral agreements) and TTL=2.

For this analysis, the total set of nodes has been ordered according to the size of the hosted data and grouped in 4 categories, with specific characteristics.

1. **N1:** Category of the set of nodes with the smallest size of their datasets, with a total amount corresponding to the first 25% of the data available in the network.
   This group accounts for 146 nodes. Each of them collects (in average) 0,31% of the total data collected through bilateral agreements, and this figure grows up to 0,48% using multilateral agreements (TTL=2). This group of nodes is the only one where the amount of data collected through multilateral agreements is bigger than the amount collected through bilateral ones.

2. **N2:** Category of the set of nodes with a size of their datasets larger than those in N1, with a total amount corresponding to the second 25% of the data available in the network.

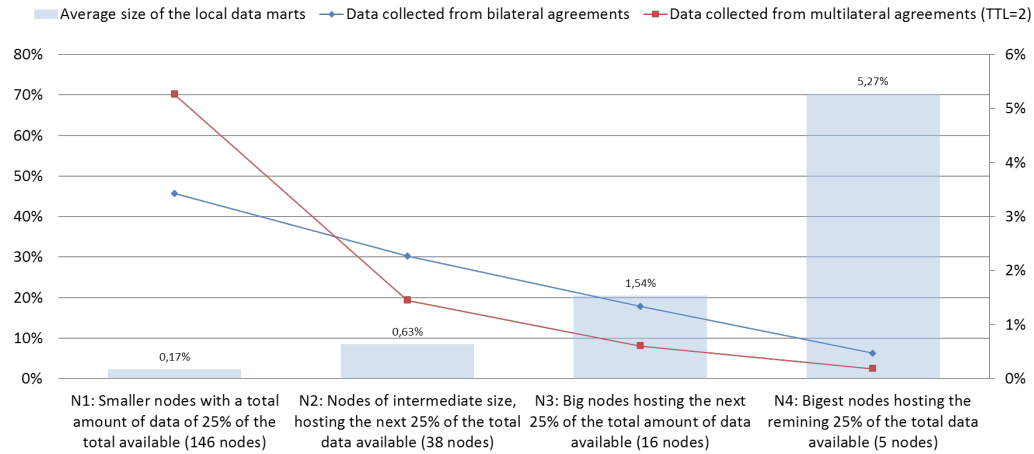

Figure 7.6: Data collected from bilateral agreements and multilateral ones (TTL=2) according to the specific characteristics of the nodes, grouped in four categories depending on the size of their local data marts.

This group accounts for 38 nodes and each of them collects (in average) 0,80% of the total data collected through bilateral agreements, and 0,51% using multilateral agreements (TTL=2).

3. **N3:** Category of the set of nodes with a size of their datasets larger than those in N2, with a total amount corresponding to the third 25% of the data available in the network. This group accounts for 38 nodes and each of them collects (in average) 1,11% of the total data collected through bilateral agreements, and 0,50% using multilateral agreements (TTL=2).

4. **N4:** Category of the set of nodes with the biggest size of their datasets, with a total amount corresponding to the last 25% of the data available in the network.
   This group accounts for only 5 nodes and each of them collects (in average) 1,25% of the total data collected through bilateral agreements, and 0,49% using multilateral agreements (TTL=2).

Part IV

Conclusions

Chapter 8

Achievements and Open Challenges

"Science never solves a problem without creating ten more"
George Bernard Shaw (1856-1950)

8.1 Executive summary of the results

The advent of the Big Data era has a strong impact in many sectors, including health. With the growing amounts of heterogeneous, disperse and worldwide distributed clinical and biomedical data repositories, novel approaches, methods and tools for their management, process and analysis are required. A significant part of the future of medicine depends on the combined analysis of this data.

Public regulations and clinical advances towards personalised medicine pushes the integration and sharing of this huge amount of data, and significant efforts from diverse research projects are building the required infrastructures towards this major scientific challenge. The MOSAIC system resulting from this Thesis proposes novel strategies to facilitate the process of data sharing and exchange through multilateral agreements.

The distributed nature of the IT infrastructure in the healthcare sector builds complex networks with evolving and dynamic topologies in terms of data available, links connecting nodes, and specific data access needs. This is an ideal framework for the deployment of IT solutions based on the Multi-Agent Systems paradigm, as MOSAIC is.

The need of a distributed process to support the achievement of multilateral agreements for data exchange is justified for the lack of global knowledge of the network topology derived from the reluctance to publish certain information in a centralised repository. The use of Agents has facilitated to model the negotiation process required by the actors of this system.

The MOSAIC protocol has been designed to support the communication and negotiation process among the MOSAIC Agents. A part from the number of cases collected from the network, additional metrics for the assessment of the protocol and its efficiency have been defined. These include the number of agreements and the number of messages of the protocol

required to achieve an agreement.

An equation to formalise the theoretical performance of MOSAIC in terms of number of agreements achieved depending on the length of the path allowed for the multilateral agreements, has been formulated and validated. This equation includes a parameter that permits to adjust the shape of the resulting curve to the specific characteristics of different MOSAIC strategies for the path selection during the network exploration.

The execution of the MOSAIC System has been simulated in a realistic scenario of a network of nodes hosting brain tumours. With the MOSAIC validation in this scenario, it has been demonstrated that multilateral agreements among a set of networked nodes of a Federated Data Warehouse increase significantly the amount of data accessible in a network compared with the amount of data that can be collected from bilateral agreements.

A number of strategies for the path selection during the network exploration has been proposed, two of them have been implemented, and with them the predicted values from the equation have been compared against the real values obtained from the MOSAIC execution, validating the accuracy of the mathematical model that has very little deviations. It has been proved that the strategy to select the path to follow during the exploration of the network has implications in the number of agreements achieved among the nodes. The two criteria tested are: i) A random selection of the path and ii) Path selection based on the Dataset size. Finally, it has been demonstrated that the total number of agreements among the nodes, achieve better marks when the path selection is based on the dataset size rather than a random selection.

8.2 Future work

The research carried out in this Thesis can be extended in the following aspects:

- **Semantic representation and Constraints enrichment**. Both datasets and constraints could be represented using OWL, as the standard for knowledge semantic representation. This would facilitate the design of complex constraints based on rules and a way for the description of the datasets, ready to be used by the communications of MOSAIC's Agents. A more natural representation of the possible constraints could be based on a boolean expression composed by a set of clauses, some of them related to the delivery of a combination of certain datasets (not only a single one) and others related to the acceptance or rejection of certain top level conditions for the data access by the user.

- **Core implementation**. The optimisation of the code to allow wider and deeper path explorations of the network in a reasonable time and the visualisation of the protocol results in a web based interface.

- **Security and privacy**. Although some security features have been introduced, the

design of the Security Layer has to be deployed in depth and different aspects related to data disclosure protection, attacks prevention, and authenticity, among others should be addressed and integrated in the protocol with the deployment of previous research in the field [103, 92] and their adaptation to this specific scenario.

Granting the identity of the nodes (authenticity), and the integrity and confidentiality of the data transfered, are the basic features provided by SSL through the asymmetric and symmetric key techniques that must be part of the core of MOSAIC. Additionally, traceability of the data can be also needed in order to identify the node to which the data has been delivered and to which the access rights have been transfered. This would allow to identify the origin of the data if it is found somewhere else of the node that was allowed to access and to use it. Integrating fingerprinting algorithms into MOSAIC would grant this feature. Moreover, when the agreement of a certain negotiation implies the exchange of data, the MOSAIC protocol has to grant that the data sent can be only accessed and used if the data received can also be accessed and used. Techniques of Fair Exchange protocols could be integrated into MOSAIC to address this security feature. Finally, monitoring the activity of the Agents may provide confidential information of the Agents and their users (how many data they are exchanging, with which nodes, how often, etc) and with this, new features to protect this privacy issues can be taken into account.

- **System deployment and evaluation in different scenarios**. It is expected that the results of the protocol will differ significantly depending on the specific scenario and characteristics of the network. Therefore, it is also planned to adapt the behaviour of the Agents to different frameworks and to identify which strategies are the best for each case and specifically the best balance between path length and branch selection wide (number of branches to explore among all possible options).

- **On-line Network Analysis**. This refers to a dynamic classifier calculated at MCC level, based on stream data mining techniques, and updated in real-time according to the dynamic behaviour of the negotiation process. This strategy would learn from the experience and the successful or failed attempts to reach a dataset after after every request.

- **Game theory**. During the network exploration the MCP launches a request to get the desired data or to solve a constraint (if it is not the first MCP in the path). A set of MCC may answer and one or a set of them have to be selected.

  On the one hand, the selection of the best MCC depends on the decision of the set of nodes or MCC-MCP pairs already involved in the agreement. All of them share the goal to complete the path and achieve the multilateral agreement with the new MCC that has to be selected. All of them want to maximise the chances to achieve a successful agreement and will share the information to get the best decision. On the other hand, the set of MCC

candidates compete among them to be selected and be part of the multilateral agreement. This can be modelled as an auction where the buyers are the set of participants at the partial path already built and the sellers are the MCC candidates. The MCC compete among them, but they could also agree among them some strategy and build some coalition in order to overcome their rivals.

This concludes the dissertation of the thesis, which compiles the most salient results achieved during the last years of exciting research, opening new opportunities for deployment and exploitation of the results and future new research projects.

Part V

Dissemination of the Results

Chapter 9

Articles and Talks in Scientific Events

"The advancement and diffusion of knowledge is the only guardian of true liberty"
James Madison (1751-1836)

9.1 The MOSAIC publications

The Development of the MOSAIC system started in 2009 and in 2010 was presented at the VI Workshop on Agents Applied in Healthcare, celebrated in the framework of the eHealth2010 CCIA conference in Casablanca (Marroco). The proceedings of this conference were published in 2012, including the first article of MOSAIC [75]. The advances of the system were presented at the *X Jornadas de Ingeniería Telemática* (JITEL 2011) on September 2011 [74].

MOSAIC was also part of the 2012 Barcelona Forum on Ph. D. Research in ICT [71] and the results of its validation, together with the performance evaluation of the protocol, were presented at the VII Workshop on Agents Applied in Healthcare in the framework of the AAMAS conference in Valencia (Spain) [72]. The analysis of the security aspects of MOSAIC produced also two publications [28, 29].

The description of the final version of the MOSAIC System was published on November 2014 in a special supplement of the Open Access JCR Journal of Translational Medicine with an impact factor of 3,99 [70]. This article with the title *"Knowledge sharing in the health scenario"* received 1.031 accesses at the Journal's web site (during 2015), and it has been already cited by two other research articles.

9.1.1 International publications

Research articles and proceeding papers from international conferences and workshops:

Lluch-Ariet, M., de la Torre, A.B., Vallverdú, F., Pegueroles-Vallés, J. *Knowledge sharing in the health scenario.* Central, B. (ed.) Systems Medicine in Chronic Diseases: COPD as a Use Case. Springer, United Kingdom (2014). doi:10.1186/1479-5876-12-S2-S8

Brugués-de-la-Torre, A., Lluch-Ariet, M., Pegueroles-Vallés, J. *Security analysis of a protocol based on multiagents systems for clinical data exchange.* Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference On, pp. 305311 (2013). doi:10.1109/CISIS.2013.56

Lluch-Ariet, M., de la Torre, A.B., Pegueroles-Vallés, J. *Performance evaluation of mosaic: A multi agent system for multilateral exchange agreements of clinical data.* Proceedings of the VII Workshop on Agents Applied in Healthcare (A2HC 2012): 4 June 2012; Valencia, pp. 719 (2012). 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)

Lluch-Ariet, M., Pegueroles-Vallés, J. *The mosaic system.* Szomszor, M., Kostkova, P. (eds.) Electronic Healthcare. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 69, pp. 275284. Springer, Berlin Heidelberg (2012). doi: 10.1007/978-3-642-23635-8 35

9.1.2 National publications

Proceeding papers from national conferences and workshops:

Brugués-de-la-Torre, A., Lluch-Ariet, M., Pegueroles-Vallés, J. *Análisis de seguridad de un protocolo de intercambio de datos clínicos basado en sistemas multiagente.* University, M. (ed.) Proceedings of the XII Spanish Meeting on Cryptology and Information Security (RECSI 2012): 4-7 September 2012; Donostia (2012). Telematics Group, Mondragon University

Lluch-Ariet, M., de la Torre, A.B. *Mosaic: Multilateral agreements for clinical data exchange.* UPC (ed.) Proceedings of the 2012 Barcelona Forum on Ph.D. Research in Information and Communication Technology: 15 October 2012; Barcelona, pp. 8788 (2012). UPC

Lluch-Ariet, M., Pegueroles-Vallés, J. *Mosaic: Un sistema de intercambio de datos clínicos con soporte para acuerdos multilaterales.* Proceedings of the X Jornadas de Ingeniería Telemática (JITEL 2011): 28-30 September 2011; Santander (Spain), p. (2011). Asociación de Telemática (ATEL)

9.2 Other publications and talks of the author

The following list details additional papers authored by the PhD candidate summarizing the background work that inspired the idea to this Thesis.

Vicente, J., Garcia-Gomez, J. M., Tortajada, S., Navarro, A. T., Howe, F., Peet, A. C., Celda, B., Lluch-Ariet, M., and Robles, M. *Ranking of brain tumour classifiers using a bayesian approach* IWANN, Sept 2009. [105]

Roset, R., Lurgi, M., Croitoru, M., Hu, B., Lluch-Ariet, M., and Lewis, P. *A visual mapping*

tool for database interoperability: the healthagents case. In Third Conceptual Structures Tool Interoperability Workshop (Toulouse, France, June 2008), CEUR-WS.org, CEUR-WS.org, pp. 44-54. [94]

Xiao, L., Vicente, J., Saez, C., Peet, A., Gibb, A., Lewis, P., Dasmahapatra, S., Croitoru, M., Gonzalez-Velez, H., Lluch-Ariet, M., and Dupplaw, D. *A security model and its application to a distributed decision support system for healthcare.* In Availability, Reliability and Security, 2008. ARES 08. Third International Conference on (March 2008), IEEE, IEEE Xplore, pp. 578-585. [112]

Xiao, L., Peet, A., Lewis, P., Dashmapatra, S., Saez, C., Croitoru, M., Vicente, J., Gonzalez-Velez, H., and Lluch-Ariet, M. *An adaptive security model for multi-agent systems and application to a clinical trials environment.* Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International (Beijing, China, jul 2007), IEEE, Ed., vol. 2, IEEE Xplore, pp. 261-268. [111]

Lluch-Ariet, M. *HealthAgents: Agent-based distributed decision support system for brain tumour diagnosis and prognosis.* Talk. BIT's Annual World Cancer Congress, Shanghai, June 2008. [68]

Lluch-Ariet, M. *Multiagent systems in clinical decision support systems.* Talk, 5th Workshop on Agents Applied in Health Care. (AAMAS) Estoril, May 2008. [69]

Bibliography

[1] R. Annicchiarico, U. Cortés, and C. Urdiales. *Agent Technology and e-Health*. Whitestein Series in Software Agent Technologies and Autonomic Computing. Birkhuser Basel, first edition, 2008.

[2] Apache. Cassandra, 2015. http://cassandra.apache.org (Last accessed: 31 December 2015).

[3] Apache. Hadoop, 2015. http://hadoop.apache.org (Last accessed: 31 December 2015).

[4] Apache. Hbase, 2015. http://hbase.apache.org (Last accessed: 31 December 2015).

[5] S. Athanasiou, D. Hladky, G. Giannopoulos, A. Garcia-Rojas, and J. Lehmann. GeoKnow: Making the web an exploratory place for geospatial knowledge. *European Research Consortium in Informatics and Mathematics News*, 2014.

[6] Axle Consortium. Advanced analytics for extremely large european databases, 2014. http://axleproject.eu/ (Last accessed: 31 December 2015).

[7] F. Bellifemine, A. Poggi, and G. Rimassa. Jade: a fipa2000 compliant agent development environment. In *AGENTS '01: Proceedings of the fifth international conference on Autonomous agents*, pages 216–217, New York, NY, USA, 2001. ACM.

[8] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Research*, 41(D1):D36–D42, 2013.

[9] S. Berger and M. Schrefl. From federated databases to a federated data warehouse system. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 394, Washington, DC, USA, 2008. IEEE Computer Society.

[10] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. Magazine, May 2001. Scientific American Magazine.

[11] Biobanckcloud Consortium. A cloud-computing platform as a service (paas) for the storage, analysis and inter-connection of biobank data, 2014. http://www.biobankcloud.eu/ (Last accessed: 31 December 2015).

[12] Biopool Consortium. Services associated to digitalised contents of tissues in biobanks across europe, 2014. http://www.biopoolproject.eu/ (Last accessed: 31 December 2015).

[13] P. Boncz, I. Fundulaki, A. Gubichev, J. Larriba-Pey, and T. Neumann. The linked data benchmark council project. *Datenbank-Spektrum*, 13(2):121–129, 2013.

[14] M. J. Boniface, T. A. Leonard, M. Surridge, S. J. Taylor, L. Finlay, and D. McCorry. Accessing patient records in virtual healthcare organisations. In *eChallenges 2005*, 2005.

[15] R. Botsman and R. Rogers. *What's Mine Is Yours: The Rise of Collaborative Consumption*. HarperBusiness, 2010.

[16] D. Brugali and K. Sycara. Towards agent oriented application frameworks. *ACM Computing Surv*, 32(1):21–27, 1998.

[17] A. Brugues De La Torre, M. Lluch-Ariet, and J. Pegueroles-Valles. Security analysis of a protocol based on multiagents systems for clinical data exchange. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on*, pages 305–311, July 2013.

[18] L. Cardoso, F. Marins, F. Portela, M. Santos, A. Abelha, and J. Machado. The next generation of interoperability agents in healthcare. *International Journal of Environmental Research and Public Health*, 11(5):5349, 2014.

[19] E. Codd, S. Codd, and C. Salley. *Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate*. Codd and Date, Inc, San Jose, USA, 1993.

[20] E. F. Codd. Is your dbms really relational?

[21] J. M. Corchado, D. I. Tapia, and J. Bajo. A multy-agent architecture for distributed services and applications. *International Journal of Innovative Computing, Information and Control*, 8:2453–2476, 04/2012 2012.

[22] Council of European Union. Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1463666129942&uri=CELEX:31995L0046.

[23] Council of European Union. Directive 2003/98/ec of the european parliament and of the council of 17 november 2003 on the re-use of public sector information, 2003. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1432196812635&uri=CELEX:32003L0098.

[24] Council of European Union. 2012/417/eu: Commission recommendation of 17 july 2012 on access to and preservation of scientific information, 2012. http://eur-lex.europa.eu/legal-content/EN/NOT/?uri=CELEX:32012H0417.

[25] Council of European Union. Directive 2013/37/eu of the european parliament and of the council of 26 june 2013 amending directive 2003/98/ec on the re-use of public sector information, 2013. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1432196579612&uri=CELEX:32013L0037.

[26] Council of European Union. Directive (eu) 2016/680 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing council framework decision 2008/977/jha, 2016. http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1463666375173&uri=CELEX:32016L0680.

[27] D. de Jonge and C. Sierra. *Trust, Negotiations and Virtual Currencies for a Sharing Economy*, pages 363–366. Springer International Publishing, 2016.

[28] A. B. de-la Torre, M. Lluch-Ariet, and J. Pegueroles-Vallés. Análisis de seguridad de un protocolo de intercambio de datos clínicos basado en sistemas multiagente. In M. University, editor, *Proceedings of the XII Spanish Meeting on Cryptology and Information Security (RECSI 2012): 4-7 September 2012; Donostia*. Telematics Group, Mondragon University, 2012.

[29] A. B. de-la Torre, M. Lluch-Ariet, and J. Pegueroles-Vallés. Security analysis of a protocol based on multiagents systems for clinical data exchange. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on*, pages 305–311, July 2013. doi:10.1109/CISIS.2013.56.

[30] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959. 10.1007/BF01386390.

[31] A. Dubovitskaya, V. Urovi, K. Aberer, and M. Schumacher. An agent framework for dynamic health data aggregation for research purposes. Istambul, Turkey, May 2015. IX Workshop on Agents Applied in Health Care, held in conjunction with AAMAS 2015.

[32] E. Check Hayden. Global alliance for genomic and clinical data sharing. web site. http://www.nature.com/news/geneticists-push-for-global-data-sharing-1.13133 (Last accessed 31 December 2015).

[33] ECRIN-ERIC. European clinical research infrastructures network, 2014. http://www.ecrin.org/ (Last accessed: 31 December 2015).

[34] EMBL - EBI and the Wellcome Trust Sanger Institute. The ensembl project, 2014. http://www.ensembl.org/ (Last accessed: 31 December 2015).

[35] European Commission. European strategy forum on research infrastructures, 2014. http://ec.europa.eu/research/infrastructures/ (Last accessed: 31 December 2015).

[36] European Molecular Biology Laboratory (EMBL-EBI). Elixir - data for life, 2013. http://www.elixir-europe.org/ (Last accessed: 31 December 2015).

[37] European Research Infrastructure Consortium. The pan-european research infrastructure for biobanking and biomolecular resources, 2014. http://bbmri-eric.eu/ (Last accessed: 31 December 2015).

[38] European Union. The european union open data portal, 1995-2015. https://open-data.europa.eu/ (Last accessed: 31 December 2015).

[39] B. F., C. G., and G. D. Interactive workflows with wade. In *Proceedings of the 21st IEEE International Conference on Collaboration Technologies and Infrastructures (WETICE 2012-ACEC track)*, pages 10–15, Skokie, Illinois, USA, 2012. IEEE Computer Society.

[40] FIPA. Foundation for intelligent physical agents. web site. http://www.fipa.org/ (Last accessed 31 December 2015).

[41] P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. Searle. Ensembl 2014. *Nucleic Acids Research*, 2013.

[42] Foundation for Intelligent Physical Agents (FIPA). Agent communication language. web site. http://www.fipa.org/repository/aclspecs.html (Last accessed 31 December 2015).

[43] GEN2PHEN Consortium. Fp7 gen2phen project, 2011. http://www.gen2phen.org/ (Last accessed: 31 December 2015).

[44] GeoKnow Consortium. Geo know - making the web an exploratory place for geospatial data, 2014. http://geoknow.eu/ (Last accessed: 31 December 2015).

[45] E. German and L. Sheremetov. Specifying interaction space components in a fipa-acl interaction framework. pages 191–208, 2008.

[46] M. Giese, D. Calvanese, P. Haase, I. Horrocks, Y. Ioannidis, H. Kllapi, M. Koubarakis, M. Lenzerini, R. Möller, M. Rodriguez-Muro, z. Özcep, R. Rosati, R. Schlatte, M. Schmidt, A. Soylu, and A. Waaler. Scalable end-user access to big data. In R. Akerkar, editor, *Big Data Computing*. CRC Press, 2013.

[47] T. Glenn. Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, (11):759–769, 2011.

[48] T. Glenn. 2014 Next Generation Sequencing (NGS) Field Guide: Overview, 2014. http://www.molecularecologist.com/next-gen-fieldguide-2014/ (Last accessed: 31 December 2015).

[49] Global Alliance for Genomics and Health. Global alliance for genomics and health, 2014. http://genomicsandhealth.org/ (Last accessed: 31 December 2015).

[50] I. Herman. *Semantic Web*. W3C. http://www.w3.org/2001/sw/ (Last accessed 31 December 2015).

[51] W. Hersh, J. Jacko, R. Greenes, J. Tan, D. Janies, P. Embi, and P. Payne. Health-care hit or miss? *Nature*, (470):327–329, 2011.

[52] HOPE. Hospital platform for e-health. web site. http://sourceforge.net/projects/telemed/ (Last accessed 31 December 2015).

[53] J. Huang, N. R. Jennings, and J. Fox. An agent-based approach to health care management. *Int. Journal of Applied Artificial Intelligence*, 9(4):401–420, 1995.

[54] IDC. Idc health insights: Bigger data for better healthcare, 2013. http://www.idc.com/prodserv/insights/health/index.jsp.

[55] IHTSDO. nternational health terminology standards development organisation, 2015. http://www.ihtsdo.org/ (Last accessed: 01-05-2015).

[56] IMPART Consortium. Intelligent management platform for advanced real-time media processes, 2014. http://impart.upf.edu/ (Last accessed: 31 December 2015).

[57] Insight Consortium. Intelligent synthesis and real-time response using massive streaming of heterogeneous data, 2014. http://www.insight-ict.eu/ (Last accessed: 31 December 2015).

[58] IQmulus Consortium. A high-volume fusion and analysis platform for geospatial point clouds, coverages and volumetric data sets, 2014. https://iqmulus.eu/ (Last accessed: 31 December 2015).

[59] D. Isern and A. Moreno. A systematic literature review of agents applied in healthcare. *Journal of Medical Systems*, 40:1–14, 2015. 10.1007/s10916-015-0376-2.

[60] D. Isern, D. Sánchez, and A. Moreno. Agents applied in health care: A review. *International Journal of Medical Informatics*, 79:145–166, 2010. 10.1016/j.ijmedinf.2010.01.003.

[61] ISO 13606. Electronic health record communication – part 1: Reference model, 2008.

[62] ISO/HL7 27931. HL7 version 3 - Reference information model, 2006.

[63] ISO/TR 20514. Electronic health record – definition, scope and context, 2005.

[64] R. J, L. AS, H. MG, and L. JE. Eric: a new governance tool for biobanking. *European Journal of Human Genetics*, (22):1055–1057, 2014.

[65] Jade. Jade - java agent development framework. web site. http://jade.tilab.com/ (Last accessed 31 December 2015).

[66] G. A. Komatsoulis, D. B. Warzela, F. W. Hartela, K. Shanbhaga, R. Chilukuric, G. Fragosoa, S. de Coronadoa, D. M. Reevesa, J. B. Hadfielda, C. Ludetb, and P. A. Covitza. cacore version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. *Journal of Biomedical Informatics*, 41:106–123, 2008. Article in Press. DOI: 10.1016/j.jbi.2007.03.009.

[67] LDBC Consortium. Linked data benchmark council, 2014. http://ldbc.eu/ and http://ldbcouncil.org/ (Last accessed: 31 December 2015).

[68] M. Lluch-Ariet. Health agents: Agent-based distributed decision support system for brain tumour diagnosis and prognosis. Talk, June 2008. BIT's Annual World Cancer Congress - 2008.

[69] M. Lluch-Ariet. Multiagent systems in clinical decision support systems. Talk, May 2008. 5th Workshop on Agents Applied in Health Care.

[70] M. Lluch-Ariet, A. de la Torre, F. Vallverdú, and J. Pegueroles-Vallés. Knowledge sharing in the health scenario. *Journal of Translational Medicine*, 12(Suppl 2):S8, 2014.

[71] M. Lluch-Ariet and A. B. de la Torre. Mosaic: Multilateral agreements for clinical data exchange. In UPC, editor, *Proceedings of the 2012 Barcelona Forum on Ph.D. Research in Information and Communication Technology: 15 October 2012; Barcelona*, pages 87–88. UPC, 2012.

[72] M. Lluch-Ariet, A. B. de la Torre, and J. Pegueroles-Vallés. Performance evaluation of mosaic: A multi agent system for multilateral exchange agreements of clinical data. In U. R. i Virgili, editor, *Proceedings of the VII Workshop on Agents Applied in Healthcare (A2HC 2012): 4 June 2012; Valencia*, pages 7–19. 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), 2012.

[73] M. Lluch-Ariet, F. Estanyol, M. Mier, C. Delgado, H. González-Vélez, T. Dalmas, M. Robles, C. Sáez, J. Vicente, S. V. Huffel, J. Luts, C. Arús, A. P. C. Silveira, M. Julià-Sapé, A. Peet, A. Gibb, Y. Sun, B. Celda, M. C. M. Bisbal, G. Valsecchi, D. Dupplaw, B. Hu,

and P. Lewis. *On the Implementation of HealthAgents : Agent-Based Brain Tumour Diagnosis*. Whitestein Series in Software Agent Technologies and Autonomic Computing. Birkhuser Basel, first edition, 2008.

[74] M. Lluch-Ariet and J. Pegueroles-Vallés. Mosaic: Un sistema de intercambio de datos clínicos con soporte para acuerdos multilaterales. In J. 2011, editor, *Proceedings of the X Jornadas de Ingeniería Telemática (JITEL 2011): 28-30 September 2011; Santander (Spain)*, page x. Asociación de Telemática (ATEL), 2011.

[75] M. Lluch-Ariet and J. Pegueroles-Vallés. The mosaic system. In M. Szomszor and P. Kostkova, editors, *Electronic Healthcare*, volume 69 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 275–284. Springer Berlin Heidelberg, Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-23635-8_35.

[76] M. Maleszka, B. Mianowska, and N. T. Nguyen. A framework for data warehouse federations building. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 2897–2902, Oct 2012.

[77] H. Mazouzi, A. E. F. Seghrouchni, and S. Haddad. Open protocol design for complex interactions in multi-agent systems. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 517–526, New York, NY, USA, 2002. ACM.

[78] E. Merelli, G. Armano, N. Cannata, F. Corradini, M. d'Inverno, A. Doms, P. Lord, A. Martin, L. Milanesi, S. Moller, M. Schroeder, and M. Luck. Agents in bioinformatics, computational and systems biology. *Brief Bioinform*, 8(1):45–59, 2007.

[79] Michael Edelstein, Centre on Global Health Security, Synergizing Global Surveillance. Overcoming barriers to data sharing in public health: A global perspective, 2015. http://www.chathamhouse.org/publication/ (Last accessed: 31 December 2015).

[80] MongoDB. Mongodb, 2015. http://www.mongodb.org (Last accessed: 31 December 2015).

[81] A. Moreno, A. Valls, D. Isern, and D. Sanchez. Applying agent technology to healthcare: The grusma experience. *Intelligent Systems, IEEE*, 21(6):63–67, Nov 2006.

[82] National Cancer Institute and National Human Genome Research Institute. The cancer genome atlas. web site, 2006-2008. http://cancergenome.nih.gov/ (Last accessed 31 December 2015).

[83] National Institutes of Health. Nih genomic data sharing policy, 2014. (Last accessed: 31 December 2015).

[84] NIH. Sequencing costs and growth of genbank (the nih genetic sequence database) and whole genome shotgun (wgs), 2014. http://www.ncbi.nlm.nih.gov/genbank/statistics (Last accessed: 28-11-2014).

[85] OpenAIRE Project. Open access infrastructure for research in europe towards 2020, 2015-2019. https://www.openaire.eu/ (Last accessed: 31 December 2015).

[86] Optique Project. Scalable end-user access to big data, 2014. http://www.optique-project.eu/ (Last accessed: 31 December 2015).

[87] S. Parsons, P. McBurney, and E. Sklar. *Reasoning about Trust Using Argumentation: A Position Paper*, pages 159–170. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[88] C. Pedrinaci, J. Cardoso, and T. Leidig. Linked usdl: A vocabulary for web-scale service trading. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, editors, *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 68–82. Springer International Publishing, 2014.

[89] Phenomenal Project. A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data, 2015. http://cordis.europa.eu/project/rcn/194953_en.htm (Last accessed: 31 December 2015).

[90] O. S. Pianykh. A practical introduction and survival guide. 2012.

[91] R. A. Poldrack and K. J. Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, 17:1510–1517, 2014.

[92] D. Rebollo-Monedero and J. Forné. How do we measure privacy. *Upgrade*, pages 53–58, 01/2010 2010.

[93] Research2Guidance. mhealth app developer economics 2014, 2014. http://www.mHealthEconomics.com (Last accessed: 31 December 2015).

[94] R. Roset, M. Lurgi, M. Croitoru, B. Hu, M. Lluch-Ariet, and P. Lewis. A visual mapping tool for database interoperability: the healthagents case. In CEUR-WS.org, editor, *Third Conceptual Structures Tool Interoperability Workshop*, pages 44–54, Tolouse, France, June 2008. CEUR-WS.org.

[95] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, R. Forsberg, Y. Gao, V. Kashyap, J. Kinoshita, J. Luciano, M. Marshall, C. Ogbuji, J. Rees, S. Stephens, G. Wong, E. Wu, D. Zaccagnini, T. Hongsermeier, E. Neumann, I. Herman, and K. Cheung. Advancing translational research with the semantic web. *BMC Bioinformatics*, 8(3), 2007.

[96] T. R. Sahama and P. R. Croll. A data warehouse architecture for clinical data warehousing. In *ACSW '07: Proceedings of the fifth Australasian symposium on ACSW frontiers*, pages 227–232, Darlinghurst, Australia, Australia, 2007. Australian Computer Society, Inc.

[97] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, editors, *The Semantic Web – ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, pages 245–260. Springer International Publishing, 2014.

[98] SemaGrow Consortium. Data intensive techniques to boost the real-time performance of global agricultural data infrastructures, 2014. http://www.semagrow.eu/ (Last accessed: 31 December 2015).

[99] C. Sierra and J. Debenham. *Building Relationships with Trust*, pages 485–507. Springer Netherlands, Dordrecht, 2013.

[100] K. A. Spackman, K. E. Campbell, and R. A. Côté. Snomed rt: A reference terminology for health care. In *J. of the American Medical Informatics Association*, pages 640–644, 1997.

[101] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. SNOMED clinical terms: Overview of the development process and project status. In *Proc AMIA Symp.*, pages 662–666, 2001.

[102] The National Human Genome Research Institute (USA). Workshop on establishing a central resource of data from genome sequencing projects, 2012. http://www.genome.gov/27552142 (Last accessed: 31 December 2015).

[103] J. Tomàs-Buliart, M. Fernández, and M. Soriano. Protection of mobile agents execution using a modified self-validating branch-based software watermarking with external sentinel. 5508:287–294, 2009.

[104] W. van Panhuis, P. Paul, C. Emerson, J. Grefenstette, R. Wilder, A. Herbst, D. Heymann, and D. Burke. A systematic review of barriers to data sharing in public health. *BMC Public Health*, 14(1):1144, 2014.

[105] J. Vicente, J. M. Garcia-Gomez, S. Tortajada, A. T. Navarro, F. Howe, A. C. Peet, B. Celda, M. Lluch-Ariet, and M. Robles. Ranking of brain tumour classifiers using a bayesian approach (icann 2009). Limassol, Cyprus, Sept 2009 2009.

[106] S. Vitabilea, V. Contib, C. Militellob, and F. Sorbello. An extended jade-s based framework for developing secure multi-agent systems. *Computer Standards and Interfaces*, 2008.

[107] W3C. Resource description framework (rdf), 2014. http://www.w3.org/RDF/ (Last accessed: 31 December 2015).

[108] A. J. Webb, G. A. Thorisson, A. J. Brookes, and on behalf of the GEN2PHEN Consortium. An informatics project and online "knowledge centre" supporting modern genotype-to-phenotype research. *Human Mutation*, 32(5):543–550, 2011.

[109] WHO. International statistical classification of diseases and related health problems - 10th revision, 2015. http://apps.who.int/classifications/icd10/browse/2015/en (Last accessed: 31 December 2015).

[110] H.-E. Wichmann, K. A. Kuhn, M. Waldenberger, D. Schmelcher, S. Schuffenhauer, T. Meitinger, S. H. R. Wurst, G. Lamla, I. Fortier, P. R. Burton, L. Peltonen, M. Perola, A. Metspalu, P. Riegman, U. Landegren, M. J. Taussig, J.-E. Litton, M. N. Fransson, J. Eder, A. Cambon-Thomsen, J. Bovenberg, G. Dagher, G.-J. van Ommen, M. Griffith, M. Yuille, and K. Zatloukal. Comprehensive catalog of european biobanks. *Nature Biotechnology*, (29):795–797, 2011.

[111] L. Xiao, A. Peet, P. Lewis, S. Dashmapatra, C. Saez, M. Croitoru, J. Vicente, H. Gonzalez-Velez, and M. Lluch-Ariet. An adaptive security model for multi-agent systems and application to a clinical trials environment. In IEEE, editor, *Computer Software and Applications Conference, 2007. COMPSAC 2007. 31st Annual International*, volume 2, pages 261–268, Beijing, China, jul 2007. IEEE Xplore.

[112] L. Xiao, J. Vicente, C. Saez, A. Peet, A. Gibb, P. Lewis, S. Dasmahapatra, M. Croitoru, H. Gonzalez-Velez, M. Lluch-Ariet, and D. Dupplaw. A security model and its application to a distributed decision support system for healthcare. In IEEE, editor, *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*, pages 578–585. IEEE Xplore, March 2008.

[113] A. J. Zaveri, J. Lehmann, S. Auer, M. Hassan, M. Sherif, and M. Martin. Publishing and interlinking the global health observatory dataset. *Semantic Web Journal*, 2013.

[114] S. Zhou, A. Zhou, X. Tao, and Y. Hu. Hierarchically distributed data warehouse. In *HPC '00: Proceedings of the The Fourth International Conference on High-Performance Computing in the Asia-Pacific Region-Volume 2*, page 848, Washington, DC, USA, 2000. IEEE Computer Society.

Network Engineering Department
Universitat Politècnica de Catalunya