POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Executive Summary of the Thesis

A metadata model for healthcare: the Health Big Data case study

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Nives Maria Migotto

Advisor: Prof. Cinzia Cappiello

Co-advisors: Prof. Pierluigi Plebani, Prof. Letizia Tanca

Academic year: 2021-2022

1. Introduction

Big data pose a challenge to the traditional data management system, as greater storage, more complex processing, and more flexible analyses are required. In the healthcare field, a considerable amount of information is generated and used in all applications, increasing together with the technological progress, including patient personal information and medical history, which are stored in electronic health records, data from imaging and laboratory examinations, data from genomics-driven experiments, and data generated by monitoring devices. Big data in healthcare can bring various benefits. It can help patients make the right decision in a timely manner. Collecting different data from different sources can help researchers and developers by improving research on new diseases, therapies and technologies. Healthcare providers may recognize high risk populations and act accordingly (i.e. propose preventive measures), enhancing patient experience [2].

New architectures were therefore developed as an answer to the need for innovative and efficient ways to handle big data. Among these new architectures, the data lake, which has been adopted in this work, represents an emerging approach as a repository that supports structured, semi-structured and unstructured data at any scale. However, data lakes, as well as the other solutions, need to include a governance layer to effectively maintain the value of the data. A fundamental tool in data governance is the data catalog. It relies on additional data describing the managed resources, called metadata. More in detail, metadata execute two main functions: on the one hand, they aid in the structure, preservation and regulation of data, on the other hand they describe the data facilitating their discovery and use.

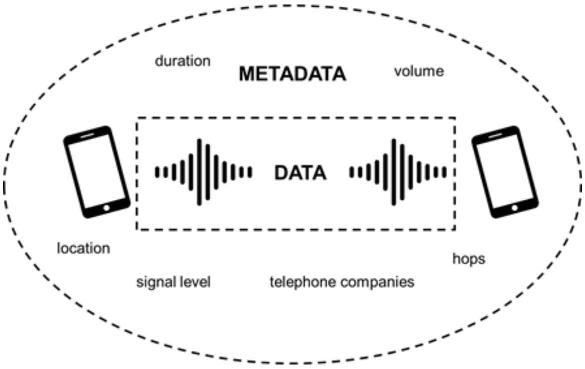

Figure 1: Data and metadata

2. Objectives

The aim of this thesis is to propose a metadata model to organize data sets related to the healthcare domain in a data catalog. In particular, this thesis has been developed in the context of the Health Big Data project, where all the IRCCSs (Istituti di Ricovero e Cura a Carattere Scientifico - institutes for treatment and research) are involved to create a common approach to share data for clinical trials.. A suitable metadata set, integrated into a data catalog, would be helpful at different levels. First, it would assist in the integration and management of the data, key point at the base of all other services. Moreover, it would enable data scientists and medical professionals to find and retrieve in a functional and timely manner the data relevant to their needs, be it to enroll patients in a clinical trial, look for previous cases relevant to correctly formulate a diagnosis, examine laboratory results, connecting exams to the patient they belong to, and so on. Metadata would make it possible, for instance, to group data based on some attribute, link related objects, and filter search results.

A feature of metadata that is scarcely considered is that they are a human construct. Metadata are designed by human beings for a particular purpose, thus the form they take strongly depends on their cause and they offer a subjective view about the objects they describe [3]. For this reason, different domains, or even different applications within the same context, require a suitable metadata set, developed specifically with those data and uses in mind. This explains why it is difficult to find a suitable metadata set in the healthcare domain. They are typically either too general, and consequently insufficient for the needs of the application at issue, or very specific to particular sub-fields, which either fall outside of the considered scope or represent only part of it.

With this in mind, the project presents a custom metadata model designed taking into account the needs and requirements specific to this project.

3. Method

The process leading to the development of the metadata model consists of two steps. The first is a thorough review of the literature, looking for classifications that could be useful for this application. Then, the project requisites were more accurately examined and some considerations were made on the metadata themselves.

3.1. Literature review

Going into more detail about the review, a number of published papers on metadata, with the addition of a few online articles from qualified sources, were surveyed, both all-round and expressly about healthcare and medical applications. Sorting through the results, the focus was on the ones proposing a classification. Among them, the papers presenting the most pertinent ones for this project were selected.

Among the most relevant metadata models, it is worth mentioning the one offered by Gilliland [4], who distinguishes between administrative (used in managing and administering resources), descriptive (used to identify and describe resources), preservation (related to the preservation of resources), technical (related to how a system functions or metadata behave), and use metadata (related to the level and type of use of resources).

Another interesting perspective, focused on the purposes of metadata in the healthcare field, is Moehrke's [5]. The categories he identified include provenance (describing where the data come from), security and privacy (used by privacy and security rules to appropriately control the data), descriptive (used to describe the clinical value, so they are expressly healthcare specific), exchange (enabling the transfer of the data), and object lifecycle metadata (describing the current lifecycle state of the data, including relationships to other data).

Analyzing the results of the survey, it was found that in the various models different criteria were used, different scopes were covered, different terms were used with the same meaning or, conversely, the same term with different meanings, and similarly identified classes overlapped. In addition, a considerable difference between general scope and healthcare-related categorizations is the appearance of numerous domain specific classes and items in the latter. All these issues can be reduced to the fact that, as stated above, each metadata schema is tailored to the specific application at hand. In fact, different applications have different requirements and focus points, which reflect in the choice of the relevant metadata. These findings are further proof that an ad hoc metadata model is necessary.

3.2. Requirements and metadata analysis

Moving to the project analysis, the focus was primarily on the context, i.e. a federation of around 50 IRCCSs, the system architecture, i.e. the data lake, the different types of data, i.e. medical records, diagnostic images and signals, and omics data, and the final uses, i.e. treatment and research, with the related requirements, e.g. high need for privacy while still exchanging data to obtain the best possible results.

As far as the considerations on metadata are concerned, the most relevant regards the so-called technical metadata, which represent the technical aspects of data that are necessary for data presentation, manipulation, and analysis. While in several papers they are regarded as a category in and of itself, we believe that they are better characterized as transversal. In fact, while each of the other classes identifies a specific topic, a 'what', technical metadata can be seen as the 'how' of all these classes. In other words, instead of having a dedicated technical metadata category, every other one would have (if necessary) a section dedicated to technical details.

Another issue regards the commonly named descriptive metadata. They are sometimes broadly used to denote anything describing data, from the content to the structure and relationships, while other times some of these aspects are regarded as individual classes. The solution that seemed more suited to resolve this matter is a hierarchical structure, in which descriptive metadata are again partitioned into subsets.

4. Metadata model

4.1. The model

The metadata model proposed in this thesis is organized around three general classes: governance, data lifecycle management and descriptive metadata. Descriptive metadata have been further divided into business/semantic, intrinsic, and inter-relationship metadata, due to the broadness of their scope.

Governance metadata cover all security and privacy policies, access rights, ownership and responsibility roles, acquisition information, data quality, data authenticity, and other legal requirements.

Data lifecycle management metadata are related to data provenance, including the source of the data, all transformations performed and existing versions, usage tracking, and information required to preserve and use the data, including technical specifications.

Descriptive metadata describe data for purposes of identification and discovery. Since this is a very comprehensive definition, they are further divided into subcategories: business/semantic, intrinsic and inter-relationship metadata.

Business/semantic metadata describe the meaning of data through descriptions, tags, indexes, attributes, etc. In addition, they include constraints and other relationships within each datasets. Their usefulness can be increased by compiling a knowledge base and employing it to annotate the data. As a result, different items referring to the same concept or connected by some semantic relationship can be connected and retrieved more easily.

Intrinsic metadata describe the characteristics of the schema and its values. They include profiling, statistics and descriptive-technical metadata. Inter-relationship metadata pertain to relationships among datasets.

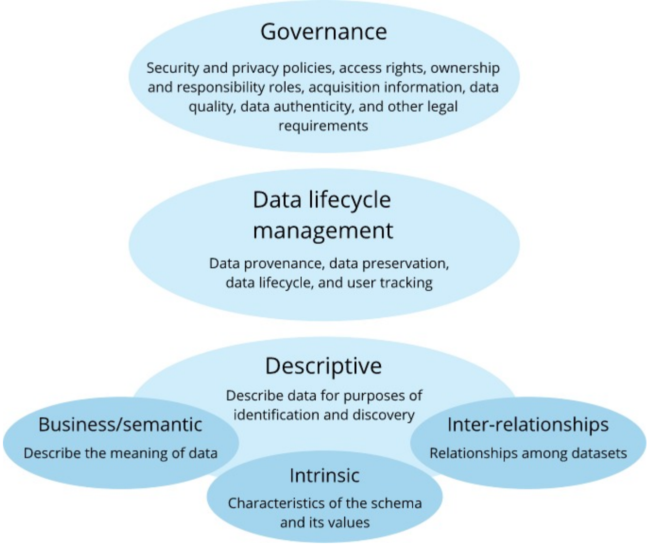
Governance
Security and privacy policies, access rights, ownership and responsibility roles, acquisition information, data quality, data authenticity, and other legal requirements
Data lifecycle management
Data provenance, data preservation, data lifecycle, and user tracking
Descriptive
Describe data for purposes of identification and discovery
Business/semantic
Describe the meaning of data
Inter-relationships
Relationships among datasets
Intrinsic
Characteristics of the schema and its values

Figure 2: Metadata model

4.2. Observations

Specific subclasses were attributed to one category or the other based on the agreed definitions, although it could be argued that a specific item or subclass might be better suited for a different category. This is inevitable, since the metadata cover more a continuous spectrum than precisely disjoint sets; as a result, some items are bound to be attributable to more than one group.

Point often overlooked, the distinction between data and metadata is not always clear. Not only that, but, on occasion, an object can be considered as both data and metadata within the same organization depending on the use and context. For instance, while examining the medical record of a patient, all patient information is undoubtedly data. On the other hand, if the object of interest is a medical image or other kinds of exam or laboratory results, the patient information becomes metadata providing additional insight into the image. This reasoning applies mostly to descriptive metadata, as the other kinds are service metadata, aimed at aiding the proper management and use of the data, and they do not belong to the same semantic domain as the data (i.e. medicine and healthcare). As a result of this dual nature, it would be useful to store data and metadata together in the same way.

4.3. Validation

In order to validate the model, a data catalog platform needed to be selected to implement the metadata. Focusing on open source solutions and taking into account the requirements of the application, 7 possible candidates were identified, Apache Atlas, Amundsen, CKAN, Kylo, Magda, Truedat, and iRODS. Among them, Apache Atlas [1] was chosen primarily for its flexibility, metadata management and overall features.

It is a metadata management and data governance tool that allows to ingest, discover, catalog, classify, and govern data from multiple data sources. It employs a metadata system that, besides offering a set of predefined metadata types, allows to create custom types so as to characterize the model as needed. A type is a definition of how a particular type of metadata objects is stored and accessed. Each type represents one or a collection of attributes that define the properties for the metadata object. A specific instance of a type is called an entity, and represents a specific metadata object in the real world. The values of an entity are the values of the attributes defined during the corresponding type definition. Seeing as these types are usually used to define technical aspects, Atlas made available what it calls business metadata, a particular type fully customizable, fit to capture business details that can help organize, search and manage metadata entities. Moreover, it is possible to dynamically create classifications and propagate them through data lineage, to better organize the data.

Regarding the validation, a demo implementation was realized using primarily types (and entities) and business metadata. In order to better categorize the entities, four new types were defined based on the different kinds of data considered in the project: patient data, image, signal, and omics. In turn, image, signal, and omics were provided with sub-types, so that each subtype can inherit the parent attributes and add specific ones. Except for the technical metadata associated to the types, the metadata model was then mapped to Atlas through the business metadata. Each of the three main categories is represented by a business metadata, and each metadata item by one of the associated attributes. Lastly, after defining the suited attributes of a few example entities, some queries were carried out. The results show that the defined metadata are indeed helpful in the identification and retrieval of the datasets of interest for a given query.

5. Conclusions

The work presented in this thesis is a step in the design and implementation process of a storage system for medical data. The main objective of the work is to determine a metadata model functional for the storage, maintenance and use of medical and healthcare related data in a federated setting. This was done taking into account existing solutions, in addition to the project needs and requirements.

The model defined in this thesis is a first proposal of a metadata set in the described context, seeing as nothing of the sort could be found in the literature. However, it still needs to be further developed. In particular, the metadata items should be better characterized, considering the specific features of the data and the system. The model should also be fully implemented, be it through Atlas or a better suited, possibly even ad hoc developed, tool.

Another important step that should be made is the identification of a minimum metadata set.

To clarify, the metadata should be classified based on their relevance and usefulness, distinguishing between necessary, recommended and optional metadata. The necessary metadata, essential to the correct functioning of the system, make up the minimum metadata set.

References

[1] Apache Software Foundation. Apache atlas - Overview.

[2] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Big data in healthcare: Challenges and opportunities. In *2015 International Conference on Cloud Technologies and Applications (CloudTech)*, pages 1–7, 2015.

[3] Richard Gartner. *Metadata. Shaping Knowledge from Antiquity to the Semantic Web.* Springer International Publishing, 2016.

[4] Anne J. Gilliland. Setting the stage. In Murtha Baca, editor, *Introduction to metadata*, page 9. Getty Research Institute, Los Angeles, CA, second edition, 2008.

[5] John Moehrke. Healthcare metadata, May 2014.