Chapter 2
Motivation and State of the Art

As we have seen in the previous chapter, data management and integration has become a major component of contemporary biomedical research. To be of any usefulness the flood of information produced by high-throughput genomic platforms (aCGH/SNP-arrays, DNA Microarrays, NGS technologies…) and high-resolution imaging platforms must not remain isolated in a sandbox, but must be integrated with all other available data about patient clinical history and lifestyle. This unified view is of paramount importance as healthcare paradigms move towards personalised medicine. A report published in *Nature* in 2013 highlighted that data storage and management requirements exceeded computational capabilities of one order of magnitude [1]. Management and integration of heterogeneous information has become an open biomedical research problem. The scientific community has proposed different solutions. So far, most of them are focused on specific subfields (i.e. functional genomics, mass spectrometry, computational neuroscience…). As multi-disciplinary collaborations become more and more pervasive in biomedical research, the limitations due to repositories focused on a single domain or a discipline must be overcome. In the next section I will introduce the concept of metadata, and discuss how they are used in information technology for data management. I will show that the concept of metadata as mere cataloguing tools has not changed as they were adopted in (biomedical) research, and that a new view on metadata must be attained as biomedicine is moving towards multi-disciplinary personalised medicine. Subsequently, I will illustrate the most common challenges a biomedical data repository must face, and then I will analyse some of the existing platforms. In the end I will address the missing aspects in the state of the art in order to envision a possible solution.

M. lzzo, *Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management*, Springer Theses,
DOI 10.1007/978-3-319-31241-5_2

2.1 Data Management and Metadata

The two most used words in publications dealing with data management are not unsurprisingly "data" and "metadata". Before proceeding further, I will clarify—as much as possible—the meaning of the latter in the context of biomedical research. Metadata are usually defined as "data about data", or more extensively "data about data and all the processes that produce, streamline and output data".[1] In the most general definition, metadata are nothing else but description of 'things'. The described things can be physical (books, individuals, items stored in warehouse) or digital objects (files or other resources). In information technology and data management systems, physical objects will be mapped to digital entities (e.g. rows in a database). As a consequence metadata will refer to data that reside in the same physical (i.e. file system) or logical unit (database, digital repository, enterprise…). As stated by David Marco in his book *Building and Managing the Metadata Repository*, [3] "when we mention metadata we are talking about knowledge". Without metadata we might not be able to correctly interpret the data; they often contain essential information that could not be otherwise retrieved or reconstructed. For instance, if you take into consideration a three-dimensional MRI scan, the 3D image will be unravelled to a one-dimensional sequence of pixel in order to be written to a file. The information of width, height and depth of the image must be stored as metadata otherwise the image cannot be reconstructed univocally. Before becoming a hot topic in collaborative research, metadata were traditionally used to catalogue items such as books, articles, and magazines in the card catalogues of libraries. Metadata management in library catalogues was introduced through the adoption of Integrated Library Management Systems. The library paradigm later shifted to the information technology domain, with the institution of Digital Libraries, which are collections of digital objects having heterogeneous nature: books, images, videos, and so on. In these environments the main use of metadata is for the information retrieval. Some examples of digital libraries are online long-term archives like arXiv.org [4] and the Internet Archive [5]. In the last thirty years, science has become more and more data-driven and collaborative, both across scientific domains and geographical regions. Metadata have been more and more extensively adopted in research, and have gained a fundamental role, as the key to communicate data between groups in collaborations.

[1] This is now the most widely accepted definition in biomedical research and information science as well. However, it was not always so. In computer science, the term metadata was originally introduced by Philip Bagley [2] to encompass both "descriptive" and "structural" metadata. The latter category, defined as "data about the containers of data", specifies how the data is stored within a (digital) system, and has been the object of the ISO11179 standard specifications. The "validation" metadata category is a subcategory of structural metadata that provide validation constraints. "Guide" metadata are descriptive metadata that help the users to find and retrieve their data. In my thesis, whenever I will speak about "structural metadata" I will use the terms "metadata schema", "schema", or "metadata model", to distinguish it from the "descriptive metadata", that I will call simply metadata.

2.2 Data Sharing in Biomedical Research: An Open Issue

The constantly increasing usage of advanced imaging and high-throughput platforms, combined with the improvements of networking and fast computing technologies, have brought up an environment where global, multi-disciplinary collaborations between geographically distributed researchers are getting increasingly common. Data repositories offer an efficient way to share informations and analysis outputs of a study from different institutions in a multi-site collaborative effort, to provide larger datasets and common resources. However, imaging techniques and NGS platforms come in many flavours, are acquired in different modalities and process data using different analysis pipelines. This is a well known problem in genomics: for instance, Illumina technologies adopt a sequencing-by-synthesis approach [6] that employs fluorescently labelled reversible-terminator nucleotide, while Ion Torrent "harnesses the power of semiconductor technology" detecting the protons (i.e. H^+ ions) released as nucleotides are incorporated during synthesis [7]. As a result Illumina and Ion Torrent sequences are usually with different processing pipelines, and increasing efforts are required for comparison and integration. If follows that data originating from different sources or processing tools are heterogeneous in data format representation and metadata content. In principle, metadata allow scientist separated in time and space (different institution, country, regulation...) to interpret in the correct way the data they refer to. In this sense, metadata allow different individuals or research groups to find a "common ground" or understanding of the data. Usually the safest way to achieve this common ground is the definition of detailed and standardised metadata. Since the advent of Functional Genomics and sequencing technologies, the concept of metadata in biomedicine is usually accompanied with the concept of annotation. An annotation is metadata (usually in the form of a comment, explanation or tag) attached to some resource or data. It refers frequently to a specific portion of the original data. A genome annotation is therefore metadata containing biologically relevant information attached to a genomic sequence. Genome annotation usually describes the whole process of identifying the location of genes and other coding or non-coding regions in a genome, and determining what is the biological function of these genes or regions.

2.3 Metadata Standardisation

The adoption of standards, with extensive and highly structured metadata is usually invoked as the "holy grail", the best solution to optimise heterogeneous data sharing, to allow efficient and effective data analysis and to avoid misinterpretation and wrong data usage. According to this view, data repositories should enforce good management practices and standardised metadata. Shared repository should actually encourage—or somehow force—scientists adopting the aforementioned practices. Gray et al. [8] explicitly state that extensive metadata and metadata standards are one

the keys to achieve scientific success in the contemporary research scenario. He maintains that standardised metadata ease data discovery and understanding of datasets by scientists and processing tools alike. In the next subsections, I will show how standardisation efforts have been carried within and without the biomedical research field. Doing this, I hope to highlight the inherent limitations of the existing—and of any foreseeable—standardisation approach.

2.3.1 Approaches to Standardisation

The earliest metadata standardisation efforts occurred outside the domain of scientific research, to handle resource management in environments such as (physical or digital) libraries, stores, and warehouses. The first standards, such as the Machine Readable Cataloguing (MARC) were established in the 1960s to describe items catalogued in libraries. One of the first, simplest and most widely accepted standards for digital metadata management is the Dublin Core Metadata Element Set (DCMES), a set of 15 metadata terms that constitute the minimum requirement to describe any (web) resource [9]. Despite its original aim of achieving omni-comprehensiveness—that could be paraphrased with the motto "One metadata standard to manage them all"—the shortcomings of the limited 15-element DCMES have raised criticism for not offering the richness and specificity required for resource description outside the web [10]. In Fig. 2.1 is shown an example of resource description using DCMES, outlining the role of its elements.

A wide variety of more specialised metadata standards arose in the past years to describe text documents (Text Encoding Initiative, TEI) and heterogeneous metadata objects stored in digital archives, such as the Metadata Enconding Trasmission Standard (METS) and the Metadata Object Description Schema (MODS). In traditional data management scenarios, metadata are always seen as a fixed product describing a given physical or virtual object, to help cataloguing, search, and retrieval. There is no strong need for metadata to mutate, adapt and evolve. The approach to data and metadata management did not change when they started to be applied in the research field, even if the scientific domain had different requirements and a more flexible and dynamic nature. In fact, research paradigms, approaches and methods change quickly over time, as goals are constantly adjusted and corrected to take up with new discoveries and techniques. Despite this, in biomedical research metadata standardisation is usually achieved adopting the same two strategies used with success in digital catalogues. The first is the adoption of minimum information sets—the so called "metadata checklists"—while the second involves the use of shared controlled vocabularies. The two approaches are strongly related to each other and overlap to some extent. I will detail them in the following subsections.

```
<metadata xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:identifier id="pub-id">
    urn:uuid:C42F6A69-5A7B-462B-B83E-CA6B90E3B484
  </dc:identifier>
  <dc:identifier id="isbn-id">urn:isbn:9780007322565</dc:identifier>
  <dc:title>The Silmarillion</dc:title>
  <dc:language>en-UK</dc:language>
  <dc:creator id="creator">J.R.R. Tolkien</dc:creator>
  <dc:contributor id="contributor01">Christopher Tolkien</dc:contributor>
  <dc:contributor id="contributor02">Guy Gavriel Kay</dc:contributor>
  <dc:date>2011-02-03T00:00:00Z</dc:date>
  <dc:description id="genre01">Mythopoeia</dc:description>
  <dc:description id="genre02">Fantasy</dc:description>
  <dc:description id="content-list">
    Ainulindalë, Valaquenta, Quenta Silmarillion,
    Akallabêth, Of the Rings of Power and the Third Age
  </dc:description>
  <dc:format>application/epub+zip</dc:format>
  <dc:publisher>HarperCollins Publishers</dc:publisher>
  <dc:rights>
    Copyright © 1977 The J.R.R.Tolkien Copyright Trust and C.R.Tolkien
  </dc:rights>
  <dc:source>urn:isbn:9780101010101</dc:source>
  <dc:subject>Morgoth</dc:subject>
  <dc:subject>Elves</dc:subject>
  <dc:subject>Silmarils</dc:subject>
  <dc:type>http://purl.org/dc/dcmitype/Text</dc:type>
  <meta property="dcterms:modified">2015-01-29T12:13:00Z</meta>
</metadata>
```

Fig. 2.1 The metadata element of an electronic version (EPUB 3.0 format [11]) of the book "The Silmarillion" [12], as specified using the DCMES. EPUB stores metadata in XML language and the Dublin Core Metadata Initiative (DCMI) elements are provided in the `dc:` namespace. For the EPUB 3.0 specification, the `identifier` (a universal unique ID (UUID), ISBN, ISSN or DOI), `title` and `language` elements are required together with the `modified` meta property. All the other elements are optional. Notice than more than one value can be provided for each field. The `source` element identifies the physical book from which the ebook was derived. The `type` element describes the nature of the document (i.e."text") using the DCMI Type Vocabulary. Two of the original 15 DCMES elements are missing in this example: `relation` and `coverage`. Further details about DCMES elements can be found at [13]

2.3.2 Minimum Information Requirements and Metadata Checklists

A minimum information standard is constituted by a set of guidelines for reporting experimental data derived by relevant method in biomedical science. These guidelines aim at ensuring that the data can be easily verified, analysed and interpreted by the scientific community. The overall goal of these efforts is to standardise the

annotation and curation processes of the experiments, providing specifications about which metadata is crucial together with the experiment's output data to make it comprehensive. One of the first efforts was the Minimum Information About a Microarray Experiment (MIAME), first outlined nearly fifteen years ago [14]. At that time, DNA microarray analysis was already widely adopted to generate gene expression data at a genomic scale. MIAME established a standard for recording and reporting microarray-based gene expression data, and it aimed at easing the development of databases and public repositories, together with standardised data processing tools. MIAME does not enforce a specific format, though formats that facilitate data querying such as spreadsheet (MAGE-TAB [15]) or XML (MAGE-ML [16]) were strongly encouraged. In its first specification, no terminology was adopted or specified to constrain the metadata values. MIAME compliant data are managed, stored and distributed by public repositories such as Array Express at EBI (UK), GEO at NCBI (US) and CIBEX at DDBJ (Japan). In the subsequent years, the scientific community specified other minimum information guidelines. Among others, noteworthy are the Minimum Information for publication of Quantitative real-time PCR Experiments (MIQE) [17] and the Minimum Information about a functional Magnetic Resonance Imaging study (MifMRI) [18]. In 2008 all these efforts were for the first time coordinated within the Minimum Information for Biological and Biomedical Investigation (MIBBI) project [19], which now provides a web-based, freely accessible resource for checklist projects, providing straightforward access to extant checklists, together with controlled data vocabularies, software tools and public databases. Not all the metadata checklist efforts are undertaken within the MIBBI consortium. The Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) was instituted to harmonise biobanking management and procedures across Europe. BBMRI developed a minimum data set for biobanks and research studies using human biospecimens [20]. This data set—the Minimum Information About BIobanking data Sharing (MIABIS)—consists of 52 attributes that—according to the proponents—provides the minimal description of the biobank's content. The harmonisation of metadata elements referring to biobanks using the MIABIS standard should improve and facilitate samples' discovery, resulting in time and cost savings and faster emergence of new scientific results.

The tremendous momentum generated by NGS technologies has also prompted in the last few years the need for novel standardised metadata checklists. In January 2014 *OMICS: A Journal of Integrative Biology* has launched a coordinated initiative for a comprehensive and flexible multi-omics (thus spanning through genomics, proteomics, metabolomics, and so on...) metadata checklist [21]. The initiative aims at enabling a thorough use of single and multi-omics datasets thanks to data harmonisation and improved visibility and accessibility.

All these standardisation efforts through checklists show a wide range of success and acceptance by the research community, as one can see making a citation search for each one of the resources mentioned above. Even the most widely accepted standards, such as MIAME, have their shortcomings that are summarised in dedicated publications [22]. An issue considered by the authors is that implementing MIAME requirements has turned out to be challenging. Specifically, since MIAME does

not provide an explicit format for representing data, no standard computer-readable format has been adopted. The MAGE-ML standard did not gain wide popularity, because of its high complexity, caused by the intricate structure of XML schemas. The much simpler MAGE-TAB format, which does not require any particular tool for viewing/editing, has gained some popularity and novel more general formats (such as ISA-TAB) tend to steer away from the XML complexity and keep the simplicity of spreadsheet formats. As a general rule, the authors observe that developing and adopting computer-readable standards is more difficult than adopting general guidelines. Simplicity is the key to success to gain consensus within the research community but this is not an easy task to achieve. A minimum set of information cannot fit all needs of a community and is not conceived to tackle the requirements of the increasingly dynamical and multidiscplinary field that is biomedical science. I will proceed to illustrate an example on this in Sect. 2.4.

2.3.3 Controlled Vocabularies: Taxonomies and Ontologies

While several minimum information standards or metadata checklists do not enforce the usage of controlled vocabularies, they frequently encourage their adoption. A *controlled vocabulary* is an authoritative list of terms that is used in indexing. Controlled vocabularies are adopted for consistent indexing (i.e. when indexing multiple documents) and do not necessary specify a structure or relationship between the terms in the list. Controlled vocabularies are a broad category which include more specialised classes: thesauri, taxonomies, and ontologies. A *taxonomy* is a controlled vocabulary with a hierarchical structure. There are relations between terms within a taxonomy, and they usually represent parent-child relationship. A *thesaurus*, is defined in the literature retrieval and the information science as a controlled vocabulary where all terms have relations to each other. There are typically three kinds of relationships: hierarchical (parent/child, as for taxonomies), associative and equivalent. They are scarcely used in life science, and their main field of application is the indexing of periodical literature. An *ontology* is a category of taxonomies with structure and specific types of relationships between terms. An ontology supplies a greater variety of relationships than the simple hierarchical associations supported by a thesaurus. Each relationship is specified in its function. Ontologies define relationships and attributes that are specific to a particular business area. They have gained a significant popularity and are extensively adopted in biomedical science. Ideally, an ontology well-fitted for research should enjoy the following good properties. It should be:

- open, so that the ontology and the body of data described in its terms should be available to the whole research community at no cost whatsoever. Being open also mean a keen receptivity towards changes driven by community debate;
- orthogonal, to provide the benefits of modularity and ensure the additivity of annotation;

- instantiated in a well-specified syntax to support algorithmic processing;
- equipped with a common system of identifiers to ensure backwards compatibility with legacy annotations as the ontology is updated.

The Open Biomedical Ontology (OBO) consortium applies the principles outlined above to the development of life science ontologies [23]. All the ontologies validated by OBO are collected and published at the OBO Foundry website. Each OBO Foundry ontology satisfies these requirements: (i) it possesses a unique identifier space, (ii) it is, or can be, expressed in formal language, (iii) it includes textual definitions for all terms, and (iv) it adopts unambiguously defined relationships according to the specifications outlined in the OBO Relation Ontology [24]. The Stanford Center for Biomedical Informatics Research of the Stanford University has developed BioPortal [25, 26], a community-driven repository for Biomedical Ontologies. The portal is equipped with a REpresentational State Transfer (REST) interface, to access ontologies and their components. Ontologies provided by the OBO foundry constitute an important core of the BioPortal ontologies. As opposed to OBI guidelines, there are few constraints on the BioPortal ontologies, provided that the ontology has some level of relevance to the domain of biomedicine and it is written in a supported format.

2.3.4 Biomedical Metadata and the Semantic Web

Ontologies provide a systematised knowledge for a specific domain, and as a consequence provide semantic content for each element and relation they define. The connection between ontologies and the so-called Semantic Web—a (mostly theoretical) extension of the Web where all the resources are described in a way that machines can understand and process to achieve inter-domain data linkage [27]—becomes apparent exploring the ontologies contained in BioPortal. Most of the over 300 ontologies stored on the BioPortal repository are either in OBO format or in the Web Ontology Language (OWL), a World Wide Web Consortium (W3C) recommendation for representing ontologies in the Semantic Web. The BioPortal dataset contains also metadata related to each ontology and mappings among terms in different ontologies. Mappings are either submitted by users or generated automatically by internal procedures of the system. Users can access the ontologies in BioPortal using the SPARQL query language or retrieving de-referenceable terms and ontology Internationalised Resource Identifiers (IRIs) [28] in Resource Description Framework (RDF) [29] format. The RDF data model is a core element to describe resources in the Semantic Web. It structures any semantic expression as a triple, consisting of a subject, a predicate and an object. A set of triples composes an RDF graph, where subject and object are the nodes and the predicate is a link defining a relationship. Nodes and links are univocally identified by a URI (or more in general, by an IRI). The RDF specification is implemented in various formats: the two recommended by the W3C are RDF-XML (the first standard format, an XML-based syntax) and Turtle,

a compact and human-friendly format. The eXtensible Markup Language (XML) is used to formalise metadata and metadata schemas. XML permits automated parsing by software tools and the addition of semantic content. Addition of semantic content to metadata has achieved a limited consensus, despite strong efforts by the W3C consortium and the RDF Working Group. There is strong debate in the developer community on the reasons behind this poor success and various indications have pointed out some drawbacks of the semantic tools like RDF/XML, NTriples, Turtle, SPARQL and so on. These tools do no provide native support for lists and generalised graph structures and they have been accused of "creating esoteric solutions to non-problems" [30]. In practice, most of the developers of data-driven applications do not need to adopt semantic web solutions for managing their when there are more "natural" and simpler solutions to address the problem. If this is true for software developers, it will be even more true for biomedical scientists and researchers who have different goals than spending time and efforts to provide semantic content to their datasets. In conclusion, semantic support should be provided when the people involved in a project feel comfortable in using it, but not enforced or tightly coupled to the metadata schema of the digital repository.

2.3.5 Established Standards for Data Exchange in Clinical Research

Numerous data models equipped with extensive metadata support have been proposed and adopted for Clinical Data Exchange. The most notable is the Health Level 7 (HL7) Clinical Document Architecture (CDA), a document markup standard that specifies the structure and the semantic content of a clinical document for the purpose of exchange [31]. It employs an XML format and can contain any form of multimedia including text, images, videos, and so on. CDA tracks administrative workflow together with clinical reports. An important aspect of CDA is the focus on document exchange rather than data sharing, and the XML format is adopted precisely for this scope. I will return on this point later on.

The Clinical Data Interchange Standards Consortium (CDISC) has produced various standards to manage both patient health care and biomedical research activities. Their Operational Data Model (ODM) is yet another XML-based schema designed to facilitate the regulatory-compliant acquisition, archiving, and interchanging of metadata and data for clinical research studies, and it is mainly focused on questionnaire-based clinical trials [32]. Moving aside from pure clinical scenarios, the CCLRC Scientific Metadata Model provides a general paradigm for storing metadata of scientific provenance. Unfortunately it does not provide a subject-centric view, that is seen as quintessential by the great majority of clinicians and biologists.

2.3.6 The XCEDE Schema

The XML-based Clinical and Experimental Data Exchange (XCEDE) schema provides an extensive metadata hierarchy for storing, describing and documenting the data produced by scientific studies [33]. Despite its omni-comprehensive objective—somehow similar to DCMES in scope, if not in the applicability domain—it has been used mainly in biomedical sciences and especially in computational neuroscience. XCEDE hierarchical structure models scientific experiments using a set of hierarchy levels; each level can be annotated with specific metadata. Structured data, such as time event-based data or clinical assessments and questionnaires can be stored directly within the XML schema. XML has been adopted to provide a standardised way to transport and interchange scientific data easing import/export procedures between heterogeneous data sources, development of specialised web services, local storage of experimental information within data collections, and creation of human and machine readable descriptions of the actual data. XCEDE version 2 has adopted reusable abstract data types, novel components to model analyses and terminologies, and is built on a hierarchical structure for defining experiments. Data associations are mapped using lists of elements linked by identifiers, thus permitting an easier integration with relational databases.

XCEDE is built on eight main components:

1. an **experimental hierarchy** of multiple levels that allows the subdivision of an experiment at different granularities, meaning that the user can omit some levels. The *project* represents the top level element and collects multiple *subjects* or *subject groups*. A *visit* represents the subject's appearance at the clinical institution or experimental site, and it may consist of several *study* elements. Each study contains one or more data-collecting *episodes*. An *acquisition* is defined as the data produced during a single episode.
2. **Resources**. XCEDE elements can associate data to resources, (heterogeneous external entities). There can be *information resources*—usually described by the 15 fields of the DCMES—that point to documents, publications or web pages, and *data resources* that work as pointers to external files containing bulk data or additional unstructured metadata.
3. **Protocols** are defined as a (possibly hierarchical) sequence of experimental steps.
4. **Structured Data** can be stored directly within the XML schema using the specific <data> tag. XCEDE natively supports internal storage of clinical assessments and events; the latter are defined as time interval annotated with metadata
5. **Analysis** is a component used to document the result of data processing steps. It consists of inputs, a list of the employed software tools/methods, and the output value(s) and/or file(s).
6. **Catalogues** are used to index and collect together data of interest. Catalogues can be nested within another catalogue to build a multi-level hierarchy.
7. The data **provenance** module monitors the origins of data and the processing steps. It allows repetition of processing workflows to test the results' replication.

8. A **terminology** component adds semantic content to the elements' content, linking them to terms within a terminology. Given the main focus on neuroscience experiments, there is support for Brain Atlases and ontologies deputed to large-scale experimental annotation.

2.4 The Shortcomings of Standards in Research Collaborations

In any real-life scenario—and biomedical research is no exception—standards take time to develop. The first issue is reaching a consensus, or "common ground" among people with different expertise. Usually this process requires numerous corrective steps, when the actors with different vocabularies negotiate a common terminology. In a multi-disciplinary collaboration the same word(s) may convey a different meaning for scientists trained in different disciplines. The second issue is due to a diversity of objectives. Many times, metadata standards do not meet the investigation purposes. Moreover, formal metadata conformant to a standard are not employed on the daily routine or on a local basis. They are not felt as a priority but as a hindrance by the researchers. A third complication is related to the multi-disciplinary nature of many biomedical studies: research in the field can no more be bound within the limited borders of a single domain. In practice, in any scientific work, well-refined metadata products do not exist, even though they would be strongly desirable (especially by the scientists themselves!!). The concept of metadata as a product works well in the settings where they were originally adopted: physical/digital libraries and archives and businesses equipped with inventories. In collaborative research, uncodified and informal knowledge plays a key role for correct data understanding. How is this usually achieved? Using incomplete, loosely structured, *ad hoc*, and mutable descriptions. The pieces of information are generated on-the-fly during communications among researchers, as soon as a common terminology is reached or restated. As evidenced in a study by Edwards et al. focused on collaborative project on different scales [34], metadata in scientific research can be envisioned as ephemeral process, rather than a fixed, enduring product. Edwards observes that metadata seen as an evolving and dynamical process enjoys some properties that make them difficult to handle.

The metadata process:

- is **fragmented**, as it involves many contributors (i.e. individuals and research groups);
- is **divergent**: often, beside the standardised efforts (and despite of them), other metadata appear. The latter are usually simpler, more widely used, and are communicated in "unconventional" ways, like emails, phone calls and face-to-face conversation, rather than data management systems and repositories;
- is **iterative**: both the metadata fields and their content is constantly changed and repaired as the common terminology shifts in time;

- is characterised by a **local-centred** focus, as for any scientist the internal usage of data and the personal goals come before long-term sharing purposes.

Therefore, metadata management requires some level of paradigm shift to reach a level of adaptability able to satisfy the researchers' needs. As long as the focus remains on fixed and highly defined metadata, there will be a strong friction and misunderstanding in data sharing and communication among scientists. Biomedical science makes no exception to this rule: there are different situations where the inherent complexity of biological and medical data cannot be captured or synthesised only with the adoption of standards.

In particular, data in omics and neuroscience—the fields of life science that mostly fall in my area of interest and that are converging faster one towards the other—are subjected to a variety of local constraints, institutional practices and regulations and specification due to the manufacturer of the analysis platforms. In these fields, a great part of metadata is generated by scientific instruments—such as DNA sequencers, CT/MRI scanners, mass spectrometers, …—and their elements are specified by the manufacturer. Different vendors adopt different strategies even if they are using similar metadata fields. A data repository must find a way to manage and possibly integrate different metadata schemas—that is the way metadata is stored and contained within the database or file system—describing the same data type (i.e. an MRI image, or a whole genome sequencing assay). Ideally, a digital repository should provide automated mapping to translate the data content and its metadata from the data source to the repository internal model. In practice, given the divergence of formats, this approach is not feasible. Neu et al. [35] have shown the inherent difficulties of heterogeneous data management and metadata standards in neuroimaging collaborative studies. Radiological images come in a plethora of formats, devised in different periods by groups of people having different goals. Formats are different: the most used are DICOM, Nifti, Analyze 7.5. In the early 1980s, the DICOM standard was established to make images independent by the scanner manufacturer. Over the last thirty years DICOM has gradually changed to keep pace with the evolution of scanner technologies. Even though a single standard was provided, each manufacturer has applied with different approaches and it has added proprietary metadata to store values that were not supported in the standard specification. DICOM provides both public and private tags to store metadata. The latter should be used by the vendors for non-standardised metadata. In practice, private tags are used by manufacturers even when a public tag is provided, if the public terminology is not consistent with the manufacturer requirements. For instance, many of the imaging modalities used in current medical research, like Diffusion Tensor Imaging (DTI), functional MRI (fMRI) and Magnetic Resonance Angiography (MRA) are not recognised by the DICOM standards, that collects all of them under the term "MR". To distinguish the three different modalities, the scanners adopt private tags and these are different among the manufacturers. Even if we consider the relatively constrained domain of Neuroimaging global consistency cannot be achieved for a number of motivations:

- radiological images are used in the clinical and research domain, with different purposes. From the point of view of the manufacturer, the clinical domain has often a higher priority and frequently a more consistent terminology than the research counterpart. An issue to consider when transferring radiological images from the clinical field to research is a consistent anonymisation with removal of all the metadata that refer to the patient identity and other sensitive information;
- research is less regulated and more dynamical, so it is less prone to adopt fixed and standardised terminologies (and this is the same point highlighted by Edwards);
- when different groups collaborate, they try to find compromises to achieve a terminological consistency;

The dynamical nature of research and the rapid evolution of technologies, require metadata that vary over time. Variations in metadata are required on a continuous basis for novel experimental acquisitions and for innovative research protocols, like those that employ pharmaceuticals never tested before, and that are not supported by the current terminology.

Researchers are frequently forced to adopt non-standard solutions during their daily activity. There is therefore the need for innovative data repositories that can adapt data models to mutating requirements, to describe novel data types or to extend existing ones. Standardisation alone cannot achieve heterogeneous data integration and an optimal sharing of information when scientists of different disciplines are collaborating in a multi-disciplinary project. As research collaborations are moving constantly to geographically distributed efforts among groups with different background the repositories must also implement adaptive metadata management tools to share data on a national and global scale. As a note, the end user should be able to extend the metadata model without resorting to computer science specialists. The repository should provide a graphical interface for data type definition and modification, where metadata fields can be added or modified, vocabularies and terminologies can be extended, and when necessary semantic content can be retrieved from an ontology to annotate the metadata fields.

2.5 Data Repositories: The State of the Art

I will now proceed examining the mostly used digital repositories and data management system for biomedical science, with a particular interest on neuroscience, Functional Genomics, and Integrated Biobanking. In the end I will draw a comparison to highlight the strengths and the deficiencies of these platforms to support a dynamical and adaptive management of metadata in a multi-disciplinary collaborative scenario.

2.5.1 XNAT

The eXtensible Neuroimaging Archive Toolkit (XNAT) platform is one of the oldest and more established data repositories for Neuroscience [36]. XNAT is an open source software suite developed by the Neuroinformatics Research Group of St. Louis, Missouri, to address and facilitate data management challenges in Neuroimaging studies. While XNAT supports mainly DICOM images and reports, it can, at least in principle, store data of different types. XNAT has automated tools to capture data from multiple sources, keeps them in a secure repository and distributes the data to authorised users. XNAT relies heavily on XML and XML Schema [37] for data representation and for other repository functions such as security management and generation of user interface content. XML Schema was chosen as it was the W3C standard specification to extend XML data formats. XNAT uses XML Schema Definition (XSD) to define data types and to generate custom components, graphical and logical content for its Presentation, Application and Data tiers. Moreover, XML is employed for security, input validation and queries. Imaging data are stored in their native format on the platform file system: a link to the file URI is stored within the database. XNAT provides neuroscientist with an ad hoc workflow for neuroimaging data acquisition and sanitising, that consists of automated data and metadata capture directly from the scanners followed by a strict quality control procedure. Non-imaging data are first put in a virtual quarantine and must be verified by a qualified operator. This wide adoption of XML schemas poses a certain number of drawbacks. Inefficient metadata storage and querying is solved saving numeric and textual fields directly within the database. Another problem that is not so easy to solve, as the authors recognise, is the poor efficiency of the database tables generated from the XSD, due to sub-optimal mapping. As a consequence of the dichotomy between XML schema and database representation, whenever a change is made to the data model it must somehow propagate to the database. This operation usually requires manual changes by an administrator and cannot be independently executed by a user without some level of computer literacy. The core data model of XNAT bears some resemblances with XCEDE—the authors of XCEDE state on their publication [33] that they have been developed to complement each other—and consists of three main data types: *Project*, *Subject* and *Experiment*. A single project is the "owner" of a Subject and Experiment entities, but it is possible to share them across projects. Experiment represents the event when new data are acquired. Experiment is an abstract model. From it derives *Subject Assessment*, and from this in turn derives *Imaging Session*. Imaging Session has three specialisations according to the scanning modalities accepted by the DICOM standard: *MR Session*, *PET Session* and *CT Session*.

XNAT relies on a three-tiered software architecture made of a PostgreSQL database back-end, a Java-based middleware tier usually deployed on an Apache Tomcat servlet container, and a web-based user interface. The XSD-to-database mapping happens as follows: each one of the XML Schema global elements is mapped to a single database table, with its sub-elements mapped to the table columns. Foreign key and complex associations (one-to-many and many-to-may) can be constructed from

the XSD. If there is more than one XML Schema additional tables are created using different namespaces to avoid overwriting. To allow interaction with analysis tools and external platforms, XNAT exposes a comprehensive RESTful Application Programming Interface (API). Recently XNAT has been equipped with a data dictionary service that allows to define relationships between data elements and taxonomical structures across the XNAT installation [38].

2.5.2 COINS

A candidate competitor of XNAT is the Collaborative Informatics and Neuroimaging Suite (COINS) project, developed at the Mind Research Network (MRN) of Albuquerque, New Mexico, USA. COINS is a data repository focused on centralisation of neuroimaging datasets from multiple studies. The focus is in building a single centralised infrastructure for neuroimaging datasets from multiple studies, rather than developing a distributed network, but the overall goal remains to maximise data sharing and reuse [39]. The authors highlight the documented benefits of a centralised approach: reducing costs, increasing the citations' rate (due to multiple cross-referencing), and the possibility of novel discovery through datasets reuse. The COINS framework backbone is constituted by a well structured taxonomy for data and data sharing. The taxonomy is twofold: on one side, it classifies the data by plurality (singleton or collection), medium (document or digital file), confidentiality (sensitive or insensitive), sense (data or metadata), source (recording of observation or derivation) and mode of acquisition (by humans or by instrument). On the other side, sharing is classified by a source entity (i.e. institution), a target entity, the possible sharing operations (intersections or unions of datasets across studies), the delivery venue (in situ or ex situ), the transfer method (through computer network, by courier or *a manu*), and the security (encryption on source, transmission, or target).

In the first attempt to categorise in rigorous way the difficulties posed by data sharing in a biomedical context, COINS developers identify five main challenges:

1. Secure Personal Health Information (PHI) management.
2. In situ versus ex situ sharing. The centralised repository helps avoiding the latter, that is copy or transfer of data outside the repository domain. All the users can log into the COINS platform and access the information in situ. This is actually a requirement that all the biomedical data repositories should satisfy.
3. The adoption of standardised metadata without undermining the extensibility to novel data types. The authors recognise that this is a major issue. They address the support for non-standard DICOM metadata—such as those describing DTI or fMRI acquisitions—with customised methods for the extraction of vendor-specific metadata fields. They stress the point that these automated procedures require continued maintenance for updates to the new scanners and technologies. COINS adopts an EAV catalogue to allow some additional flexibility for data types not natively supported by the system.

4. Intuitive Ease of Use (IEU): it is emphasised by the authors, and rightly so, that new users should require a very small effort to be productive when using the repository. Otherwise, they will soon stop using it, and they will look for easier and more profitable ways to share their datasets.
5. Expose a uniform, friendly, powerful query interface
6. Check data provenance and, when required, modify/correct metadata.

Similarly to the previously exposed platforms, COINS is developed using the established Java-based software technologies and relies on PostgreSQL. It currently supports the following data types: MRI, EEG, MEG, genetic data, neuropsychological and clinical assessments. The first four types are collectively labelled NeuroImaging Data (NID), while the last two are labelled as Neuro-clinical Assessments (NA). NA data and ND metadata are stored on the database, while the bulk ND data are stored on a file system. The COINS platform is built of five main modules, that I briefly describe. Users access the systems via *web portal*. Different studies can have their own web portal, reducing security risks; moreover, portals do not store any PHI or identifier. The *assessment manager* allows dual-entry conflict resolutions for NA that are entered by humans, and might require manual checking and validation. Data is submitted via web form. Only free-text and drop-down options are allowed, which I personally feel as an over-exemplification, if researchers would like to store information different from NA. Data can also be collected from a *tablet-based client* and sent to the COINS server using a web service API. NID upload is handled by a customised DICOM receiver based on the dcm4che Java library. At the time of the referenced publication, no other formats than DICOM were explicitly supported for data upload. A graphical *query builder* module provides an intuitive interface, that is based on a query-by-example approach, without requiring knowledge of database structure or table associations. Authorised users can query data from multiple studies. Frequently used queries can be saved and reused later. The most important module of COINS is the *Medical Imaging Computer Information System* (MICIS), devoted to studies, subjects and scans management. MICIS supports definition of custom subject types, for instance to distinguish between patients and controls. It is also equipped with a mechanism to unlink PHI from an entire study when the study expires. As of February 2015, COINS installation at MNR managed nearly 31,000 subjects from 558 studies, with more than 38,000 scan sessions and over 4,90,000 neuropsychological assessments [40]. In the end, the system main characteristics are the adoption of a centralised infrastructure to avoid ex situ sharing and wide usage of EAV catalogues to handle data extensibility without using explicit schemas in XML or other formats.

2.5.3 CARMEN

The Code Analysis Repository and Modelling for e-Neuroscience (CARMEN) system has been developed in the UK to provide a web-based portal platform to share and exploit datasets, analyses, code, and expertise in neuroscience. It provides four

main types of assets: data, metadata, services and workflows. The authors state that "data and metadata are currently structured for neurophisiology, but the mechanisms for data and metadata management are generic and hence the platform is applicable for any science domain", a somewhat bold assertion that is neither proved nor disproved in the publication [41]. The CARMEN system employs a data format and schema that has been agreed by the project collaborators. This might prove a strong limiting factor for the data model extensibility, since it is not possible to modify the model directly without accessing and operating on the source code. In CARMEN, users can generate pre-populated templates to ease the metadata entry procedures. The templates will automatically populate the data entry forms, and the user will just have to update the fields that change during the experimental protocol. When the system was originally published in 2011, metadata upload from XML files was considered as a future possibility but was not currently supported by the system. Despite its lack of flexibility, a noteworthy feature of CARMEN is the adoption of Storage Resource Broker (SRB), a file virtualisation system for distributed file storage. None of the other repositories presented in this section is currently integrated with a distributed file management system. CARMEN also offers tools for data analysis and workflow configuration. Processing applications are made available in an interactive Software as a Service (SaaS) model for end users, thanks to a Java wrapper module that handles command-line tools written in a variety of languages.

2.5.4 XTENS

The eXTENsible platform for biomedical Science (XTENS) digital repository was originally developed by Corradi et al. at the Department of Informatics, Bioengineering, Robotics, and Systems Engineering (DIBRIS) of the University of Genoa to support integrated research in neuroscience, with a particular focus on Neuroimaging [42]. Its data management paradigm was designed to handle a various range of situations and environments in biomedical research, and already incorporates a basic sample management system, a feature not yet supported by any other repository examined in this survey. XTENS allows the generation of several different data types according to structured schemas. In this respect, XTENS shows some point of contact with the XCEDE data model: data types are described by an XML metadata schema associated to XSD and XSL files to respectively validate its structure and define its representation [43]. The repository can be configured to store the metadata totally or partially in the database. The metadata are stored as XML descriptions inside the data table, to display the data in a rapid and dynamic way using XSL Transformations [44], and as records of specific metadata tables, to perform complex queries in an easier way. The most striking difference with XCEDE, and all the platform I have reviewed so far, is that XTENS abstracts Experiments, Studies, Visits, Episodes, and Data Acquisitions using a taxonomic model, built of two entities: *Process* and *Event*. An event can be any 'atomic' operation that is performed on patients or samples, or any processing of data or everything else related to the XTENS repository

administration and management. A process is defined as a group of sequential events or sub-processes related to an activity, allowing the creation of a flexible and yet hierarchical structure. A data instance in XTENS is defined as the output of an Event. The main innovative point of XTENS is that the extension of the data model through the definition of new data types can be easily performed by users through an intuitive graphical interface.

The XTENS repository architecture consists of:

1. A web portal, that provides a client interface and allows users to access and to manage database requests. The XTENS portal is a Java Server Pages (JSP) and servlet application deployed on Tomcat. To better enhance user experience and interactivity, various components are designed using Asynchronous JavaScript and XML (AJAX) programming technique. Client and server exchange messages using JSON through JSON-RPC protocol whenever possible.
2. A MySQL relational database. Database access from the web application is managed with MyBatis, a persistence framework that automates mapping between SQL databases and Java objects. The MyBatis persistence layer permits to adopt, if required, a different SQL RDMS (PostgreSQL, Oracle,. . .) with moderate effort. The database contains all the information about projects, patients, data and everything related to the repository management (users, groups and accesses).
3. A data grid storage element, which contains all the files associated to registered data instances. The data grid of choice is the integrated Rule-Oriented Data System (iRODS) middleware. SRB, the distributed file system adopted by CARMEN, was a precursor of iRODS. iRODS possesses an internal metadata catalogue and the administrator can configure XTENS to store metadata both on the internal database and on the grid storage metadata catalogue, or only on one of the two systems.

The users access the system using an existing LDAP or database account available on the server. Each user is associated to Access Control Lists in order to guarantee security and auditing. The access is via web browser without any client installation and in a secure way through the HTTPS protocol. Authentication and access-control is managed using the Spring Security framework. XTENS addresses possible security and privacy policies regarding the access to proprietary data and sensitive clinical data. This is achieved by a thorough customisation of user permissions, defined by administrator-defined entities called *functions*. Authenticated users are allowed to view, insert, modify and retrieve data according to the set of functions enabled for their own group. Administrators can define groups of users associated with different access permission to the application pages and functions.

2.5.5 SIMBioMS

The System for Informative Management in BioMedical Studies (SIMBioMS) was probably the first open-source platform designed to integrate phenotype information with genomic data produced by high-throughput profiling technologies. SIMBioMS

authors state that the driving force behind their development effort was the lack of dedicated system for Collaborative Projects that could provide support both for subject and biological sample management. SIMBioMS was originally developed in the framework of a multi-site project. Afterwards it proved to be sufficiently customisable and scalable to be adapted for other research collaborations focused on population genomics [45]. SIMBioMS supports data-entry via graphical forms, and provides facilities for data import and export. The platform can be configured to satisfy the MIBBI requirements, while data can be exported according to MAGE-TAB, ISA-TAB and customised XML and tab-delimited format. A query-by-form interface provides content exploration and report building. SIMBioMS has a modular structure and consists of two main components, that can be installed separately:

- The **Sample Information Management System (SIMS)**, stores and manages phenotypic, environmental and technical information about the collected samples. It provides four main data types: Patient, Visit, Sample and Aliquot. A patient can undergo many visits and have many samples stored within the system. One or more aliquots may be extracted from each sample for analysis purpose.
- The **Assay Data Information Management System (AIMS)** handles the experimental output for a variety of high-throughput technologies. When the platform was originally published in 2009, no built-in support was yet available for NGS. AIMS provides a hierarchical structure where an Experiment contains multiple Studies consisting of many Assay. The Assay entity provide the link with the SIMS module: an aliquot is used for one or more assays.

From the technological point of view, SIMBioMS is a Java-based application running on Tomcat and supported by a PostgreSQL database. The communication between the application classes and the database entities is handled by Hibernate [46], a popular Object-Relational Mapper (ORM). As it can be seen from the SIMS configuration guide available on the internet [47], a modification to the data model—like adding new metadata fields, or changing the names of the existing ones—require modifications to the Hibernate mapping files and to various XML configuration files of the application. These operations require at least a programmer or software installer with some previous experience, and does not provide a full control of the data model to the end-user (i.e. the scientist). Metadata and controlled vocabularies can be imported in the system using once again XML files, but the SIMBioMS developers advise the reader that this is an error-prone procedure and must be undertaken with great care.

2.5.6 openBIS

The Open Biology Information System (openBIS) software suite [48] is a data repository tailored for long-term collaborative projects adopting cutting-edge technologies, where migrations of data and support for new data models are frequently needed. It has been developed to support data management in systems biology since the acquisition from a source—such as a microscope, a mass spectrometer, a sequencer…—to

the publication. Similarly to XNAT and SIMBioMS, openBIS provides a canonical hierarchy for Biological data management with four main entities: *Project*, *Experiment*, *Sample* and *Data*. There is no explicit support for patient personal data and electronic health records, and no issues of anonymisation are tackled in the publication. openBIS's area of competence is systems biology and a distinct integration system is needed to integrate clinical information with the high-throughput omics outputs. Another drawback is the tight coupling between the main data types, especially between Experiment and Sample. An experiment can contain one or more samples, but the reverse is not true. Therefore, there is no support many-to-many association between samples and experiments,that limits the applicability if compared to SIMBioMS, where a sample could be fragmented in many aliquots assigned to one or more assays. There is no separation between the sample management domain and the analysis domain, as it was the case with the SIMS and AIMS modules of SIMBioMS. Nonetheless, openBIS shows a lot of nice features that were missing in the other platforms. Dedicated mechanisms for data upload, metadata annotation, and flexible querying have been developed for the most important fields of System Biology research: NGS, quantitative imaging, and mass spectrometry for proteomics and metabolomics. The API has been designed using a "loose coupling" approach, to expose a unified façade to external programs. Metadata are managed separately from bulk data to ease scalability issue. Metadata are stored within the relational database located on the so-called *Application Server*, while the bulk data is stored in one or more file systems managed by *Data Store Servers*. In this way, bandwidth consuming operations like file uploads are not performed on the system that provides metadata management, and that might be used frequently for searches and retrieval. In openBIS, datasets are immutable: once uploaded, they cannot be modified any more. If a dataset must be modified or new data are derived from it, the novel or updated data will be saved as a child dataset for internal consistency and information traceability. The data content is separated from its representation. If a dataset possesses different representations, the user can create multiple datasets within a single dataset container: openBIS will show the different representations as a single entity, making the duplicates transparent for the user. Data upload for small and medium files is handled through a web-based drag and drop interface: for the supported data types metadata is automatically extracted and saved on the database using dedicated Extraction, Transform and Load (ETL) procedures. Metadata can be exported as spreadsheets, and users can download bulk data directly from the web interface of using the command line. All the services offered by openBIS are available to external applications through a REST API, to allow third-party data retrieval, analysis, and visualisation.

The underlying openBIS data model divides metadata in three different categories:

- **Structured metadata**: these represent custom properties and annotations and are stored within the database as single fields. Each structured metadata element belongs to a *property type*, that determines whether it is a textual field, a number, an email, a hyperlink, a date or a constrained value selected from some terminology or controlled vocabulary.

- **Semi-structured metadata** can be stored as an XML schema directly on the database. There is no mention of tools for building or formatting these schemas, so my guess it that you have to provide them already well structured and they will be validated against an XSD.
- **Unstructured metadata** such as free-text can be provided as a file attachment and associated directly to the Project, Experiment or Sample entity.

An outline of openBIS data model is shown in Fig. 2.2. Structured and semi-structured metadata are stored using the Entity–Attribute–Value (EAV) model, a paradigm that we will see often used to build up metadata catalogues. The user can extend the data model creating new Property Types and attaching them to any if the four hierarchical components of openBIS. More details about the EAV model can be found in Sect. 2.6.

The software stack of openBIS, built on Java technology, is mostly similar to the other platform exposed in this section. It relies on a PostgreSQL database for the Application Server domain, and a file system with support for segmented and distributed storage for the Data Storage Server domain. Queries that return a large number of rows are optimised to reduce latency during data retrieval. openBIS adopts a classical three-tier architecture with presentation (WebGUI/API), domain (business objects for internal data processing) and data (Data-Access-Object pattern) layers. While the first two layers are public and accessible respectively by humans and machines, the third is responsible for the Create–Retrieve–Update–Delete (CRUD) operations on the database and is kept private. The openBIS group has additionally developed two separate applications for large data transfer. The first one, the CISD File EXchanger (CIFEX), permits web-based data upload of files larger than 2 GiB (that is the current limit for the HTTP protocol), while the other, Datamover, manages secure transfer on unreliable connection and limited storage space. Web services are

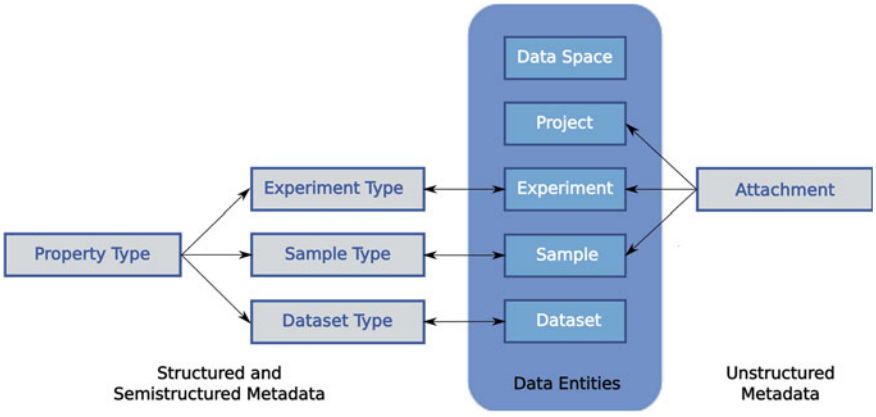

Fig. 2.2 openBIS data model. Data entities are built on a five-level taxonomy. Metadata can be either unstructured—i.e. file attachments—or (semi-)structured. The latter are described by a property type object

provided through a JSON-RPC interface, a lightweight protocol for data transmission across the internet.

2.5.7 i2b2

The Informatics for Integrating Biology and the Bedside (i2b2) platform was designed to provide clinical researchers with the tools required to integrate medical records and biomedical research data in the genomic age [49]. The i2b2 architecture is built on server-side modules, called cells, that communicate each other using web services. A set of connected cells constitute a i2b2 "hive", that can be thus extended according to custom requirements. i2b2 was designed to satisfy at least two requirements (i) find cohorts of patients that can be of interest for subsequent investigations, and (ii) use the information provided by medical records to mine the phenotype of the identified subjects in support of omics or environmental studies. The cohort of patient is selected directly from the institutional databases (e.g. clinical and biobank databases), preserving the privacy of personal information, and is copied in a project-specific data mart. A data mart is the access layer of a data warehouse that is used to expose data to the users. Data marts, like data warehouses, are read-only. i2b2 supports communication with commercial databases (Oracle and Microsoft SQL Server) to extract, transform, and load (ETL) clinical data to the data mart. The user accesses an i2b2 system from the *project management cell*, which handles authentication and authorisation. Using the i2b2 web client the authorised user can access the *data repository cell* and build up from the graphical interface ad hoc queries that are run throughout the institution (or enterprise) databases. The refined ad hoc queries will extract the data set to populate a new data mart. It is possible to select specifically which part of the data must be copied to the data mart, and specify dedicated privileges for other users. Personal data are either kept on a separate repository or encrypted within the data mart.

i2b2 data marts are based on the "star schema" design [50] of data warehousing. The schema consists of four tables:

- **observation_fact**: contains all the observations about a patient. It contains also all the value objects associated with the observation. Each value object must be a basic type (numeric, text, date,…) so composite values are stores in multiple rows. This table can be queried using and EAV approach. The observation_fact refers to the other four tables of the star schema.
- **patient_dimension**: contains the subjects' details, one patient per row.
- **visit_dimension**: describes the periods of time during which the observations were recorded.
- **concept_dimension**: contains controlled vocabulary terms that map the original codes that were used to specify the observation. The i2b2 internal vocabulary allows hierarchical grouping specifying names with a Unix-like directory approach. For instance the `/tumour` group may contain the subgroups

`/tumour/carcinoma` and `/tumour/sarcoma`. This approach allows efficient queries based on pattern matching.
- **observer_dimension**: describes the operator or the mechanical instrument that performed the recording of the observation

Overall, i2b2 collects the data of a biomedical enterprise or institution, usually located in multiple data sources, to be integrated in a small set of tables. The main implementation of i2b2 at Partners HealthCare, as of 2009, contained 1.2 billion observations from 4.6 million subjects, Observations included diagnoses, medications, procedures, and test results (including genomic test results).

i2b2 has been extended and used as a framework to integrate data in various collaborative clinical and research projects, such as the Onco-i2b2 platform [51] and the self-scaling chronic disease registry (i2b2-SSR) [52]. These implementations allow fine grained control over data integrated for sharing purposes.

2.6 Data Repositories Comparison

Confronting all the digital repositories and platforms that I have exposed in the previous section, it is possible to draw comparisons on at least three different grounds:

1. the underlying schema language/format;
2. the adaptability and ease of configurability of the metadata model;
3. the scaling properties of the system.

Concerning point 1, nearly all the metadata schemas that we have seen are structured using XML. XCEDE is written in XML, and so are the data models of XNAT and XTENS; CARMEN and openBIS have elected XML as their format of choice to storage of semi-structured metadata. In clinical data management and biomedical research XML is definitively popular. One of the main reasons for the wide adoption of XML as data-exchange format is that HL7 version 3 messages are composed in XML, and accordingly, HL7 CDA version 2 adopts the same specification. But as explicitly stated in the Release 2 reference article [31], the objective of CDA is to specify "the structure and semantics of a clinical *document* (such as a discharge summary or a progress note) for the purpose of exchange" (emphasis mine). CDA is designed to exchange documents, not data. Various software developers—Douglas Crockford [53], most notably—have pointed out that XML has document-oriented syntax poorly suited for data-oriented objects. While the distinction between the two terms—data versus document—stretches very thin and is much open to debate, from the information technology point of view I can see two requirement that a data-oriented model should satisfy: (i) allow easy mapping to object-oriented languages and (ii) be written in a language that is native in commonly used databases. With the latter I mean that the format should be fast to search and that speed performances should scale-up well with the database dimensions. Both of these statements are not true for XML. Firstly, it was not conceived to be object-oriented and requires some

level of manipulation and mapping, named data binding, to obtain a business object from an XML document tree. Many binding tools have been developed for Java, which is the language adopted by many of the biomedical repositories: the most popular are JAXB, JiBX, and XMLBeans; C++ offers CodeSynthesis XSD. With XML, data must be put within some document structure and this can be complicated with elements that can be nested, attributes that cannot be nested and complex types with their own peculiarities. When trying to develop a data model compliant with modern object-oriented languages, a software engineer should evaluate other solutions besides XML. Secondly, XML is extremely slow to search. Being a text-based format, usually more information (i.e. bits) is required to store it rather than if it were a cell value. There exists some XML-based databases—most notably BaseX, eXist and MarkLogic Server—that are optimised to use languages for XML query and navigation like XQuery and XPath/EXPath [54]. BaseX is the only open-source solution that offers extended language features without recurring to proprietary extensions. In practice though, query performance and scalability is limited, and in many cases XML is stored as text field in a relational database. Such is the case in all the data repositories I have examined. Thirdly, XML is extremely verbose, with both opening and closing tags for each element, and not conceived to be read by humans. Overall, XML works well as an exchange format for transferring data across applications adopting the same data structure, but it is too rigid too allow the flexibility to model the fluid metadata process that could simplify data sharing in research collaborations.

The second point is to some degree related to the first: the more rigid is the adopted language/format, the less adaptable will be the data model. This is the case of XNAT. Existing data types can be extended adding new fields whose values will be stored in a EAV catalogue. The operation of creating a new data type—like a new observational or clinical assessment type—requires first the construction of a new XML document that is likely to daunt the great majority of clinical users. The procedure can be eased with the help of one the many available graphical editors like Liquid XML or Eclipse Vex. Even so, once you have made a new model you have to run an update script, update the database, redeploy the XNAT application and setup XNAT security to allow access to the newly defined data types. All these operations require an administrator of the repository with good computer literacy, and this will likely take the control of the data model away from the scientists. SIMBioMS suffers from similar limitations due to the Hibernate ORM mapping files. COINS and openBIS consent a greater level of extensibility resorting to the EAV paradigm. In openBIS new metadata fields are created a Property Types and be attached to existing entities, while COINS allows the creation of customised clinical assessments to complement the neuroimaging scans. However, it does not support user-configurable fields for all the neuroimaging data types (only for MR and MEG), and no explicit creation of new data types is available to integrate other data sources. While the EAV approach provides a useful tool for defining new metadata fields, it does not explicitly offer a method to construct new data schemas. The structure of the schema must be reconstructed from the associations declared in the database tables. Compared with the other systems, XTENS aims at finding a solution to generate new data types giving full control to the end user (i.e. the scientist). It provides

non-IT users with an experience friendlier than XNAT, removing all the burdensome compilation and re-deploy steps: the user builds up your schema using a graphical web form and once submitted the XML document is automatically generated and ready for use. The administrator has only to configure the permissions to the new data type for the authorised user groups, and this is done from graphical interface as well. If a the new data type has a large set of metadata fields, the construction from web form will be an overlong and tiresome procedure. In this circumstance, it would be necessary to resort to an XML graphical editor. A good approach here is keeping the driving principle of XTENS flexibility, while providing a different, more object-oriented language for composing the metadata schema.

The issue of scalability (third point) is twofold: on the one side we must consider the database and on the other the file system for bulk data storage. Concerning the database, here the main issue is the storage of metadata: besides the adoption of XML for semi-structured metadata, nearly all the systems I have presented either put all the structured metadata in an EAV catalogue or adopt a mixed model, where dedicated tables for widely used data types (e.g. subjects, samples and/or MRI scans) exist along with an EAV representation for all the remaining metadata. COINS, SIMBioMS, and openBIS fall in the first category, while XTENS and XNAT belong to the second. EAV is very attractive because it allows a higher flexibility and requires a small level of database modelling to get it to work: in the simplest implementation, just three tables are required: one for the Entities (that in our case are the data instances), one for the Attributes (the metadata fields), and one for the attribute Values (metadata fields' values and possibly measure units). EAV presents a number of challenges that could hinder with the scalability of the system, in particular a sensitive degradation of performance: if the catalogue grows beyond a certain size, it will reach a point where the efficiency of data retrieval and manipulation will hit a critical low. At that point, the database manager or the application developer has very few choices to solve the issue. It won't be possible to add table indexes, because the table has no specific columns for each attribute as it is the case in a standard relational representation. In general, EAV is less efficient in data retrieval than "conventional" relational schema. As noted in previous studies dedicated to Clinical Databases, attribute-centred searches, where the query criterion is based on the value of a particular attribute, are most likely to show impaired performance [55]. This performance degradation is especially noticeable when query criteria combine one or more simple conditions using boolean operators. The cause for the potential performance degradation is the conversion from the relatively fast AND, OR, and NOT operations that are used when operating on relational schema tables to the sensitively slower set-based equivalents (intersection, union, and difference) for EAV tables. This study, even though fairly ancient, provides us some interesting insight on the limitations of EAV. They found that EAV query performance was three to twelve times slower than the relational schema equivalent, with search speed decreasing as the query complexity increased. The EAV structure consumed approximately for times the size of a conventional schema. As stated by Nadkarni et al. [56] the EAV model is useful for specific scenarios:

1. When dealing with metadata fields that are both numerous and sparse. A metadata field is sparse if it is present only in a small number of data instances. The case for this could be a clinical data repository handling different specialities;
2. When, even if metadata fields are not sparse, the number of different data types is large, and the number of instances for the data types is reasonably small;
3. When dealing with so-called "hybrid" data types: where some fields are sparse and some are not. In this case, even if it might be suitable, the EAV model represents a sub-optimal solution.

As we can see, there are many scenarios that fall outside the three categories outline above. As high-throughput omics technologies become more affordable and extensively used, there will be a flood of metadata—such as the variant calls from a whole genome sequencing analysis—that are not handy to wield and fast to search if stored using EAV approaches. To assess file system scalability there are two main options: (i) a distributed file system managed by a resource broker or a data grid middleware, (ii) taking advantage of a cloud-based storage. While the second is an appealing solution, it not always feasible, especially when the files contain sensitive information and the Institution regulations do not allow storing data in third-party companies' servers.

In conclusion, there is the need for a data model, based on a more flexible, object-oriented and possibly with better performances than the structures that are currently adopted, mostly XML and EAV schemas. The data model should handle in an uniform yet flexible way the wide range of heterogeneous data that are found in the current biomedical research scenario: clinical records, omics and imaging data, and biological specimens. Ideally speaking, from the developer point of view, the model should adopt a format/language that is both natural to the database where it will be stored and, in perspective, to the applications that are going to use it. I have chosen the JavaScript Object Notation (JSON) [57] format as a serious candidate to build the model, because it satisfies all the conditions I have required, and it has become a valid alternative to XML as a data exchange model for many data-driven application on the web.

References

1. Marx, V.: Biology: the big challenges of big data. Nature **498**(7453), 255–260 (2013)
2. Bagley, P.R.: Extension of programming language concepts. Technical report, DTIC Document (1968)
3. Marco, D.: Building and managing the meta data repository. A full lifecycle guide. Wiley, New York (2000)
4. arXiv.org e-print archive. arXiv.org. Accessed 18 Jan 2015
5. Internet archive: digital library of free books, movies, music & wayback machine. https://www.archive.org/ (2015). Accessed 18 Jan 2015
6. Fuller, C.W., Middendorf, L.R., Benner, S.A., Church, G.M., Harris, T., Huang, X., Jovanovich, S.B., Nelson, J.R., Schloss, J.A., Schwartz, D.C., et al.: The challenges of sequencing by synthesis. Nat Biotechnol. **27**(11), 1013–1023 (2009)

7. Merriman, B., Torrent, I., Rothberg, J.M., R & D Team, et al.: Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis **33**(23), 3397–3417 (2012)
8. Gray, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., Heber, G.: Scientific data management in the coming decade. ACM SIGMOD Rec. **34**(4), 34–41 (2005)
9. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. Internet Eng. Task Force RFC **2413**(222), 132 (1998)
10. Harper, C.: Dublin core metadata initiative: beyond the element set. Inf. Stand. Q. **22**(1), 20–28 (2010)
11. EPUB Publications 3.0.1. http://www.idpf.org/epub/301/spec/epub-publications.html (2014). Accessed 29 Jan 2015
12. Tolkien, J.R.R.: The Silmarillion. Random House LLC, New York (1979)
13. Dublin Core Metadata Element Set, Version 1.1. http://www.dublincore.org/documents/dces/ (2012). Accessed 29 Jan 2015
14. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al.: Minimum information about a microarray experiment (MIAME) toward standards for microarray data. Nat. Genet. **29**(4), 365–371 (2001)
15. Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, E., Irizarry, R.A., Liu, J., Maier, D.S., Miller, M., et al.: A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinform. **7**(1), 489 (2006)
16. Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., et al.: Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. **3**(3), research0046 (2002)
17. Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al.: The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin. Chem. **55**(4), 611–622 (2009)
18. Poldrack, R.A., Fletcher, P.C., Henson, R.N., Worsley, K.J., Brett, M., Nichols, T.E.: Guidelines for reporting an fMRI study. Neuroimage **40**(2), 409–414 (2008)
19. Taylor, C.F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.-A., Bogue, M., Booth, T., et al.: Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat. Biotechnol. **26**(8), 889–896 (2008)
20. Norlin, L., Fransson, M.N., Eriksson, M., Merino-Martinez, R., Anderberg, M., Kurtovic, S., Litton, J.-E.: A minimum data set for sharing biobank samples, information, and data: MIABIS. Biopreservation and biobanking **10**(4), 343–348 (2012)
21. Kolker, E., Özdemir, V., Martens, L., Hancock, W., Anderson, G., Anderson, N., Aynacioglu, S., Baranova, A., Campagna, S.R., Chen, R., et al.: Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. OMICS: J. Integr. Biol. **18**(1), 10–14 (2014)
22. Brazma, A.: Minimum information about a microarray experiment (MIAME)-successes, failures, challenges. Sci. World J. **9**, 420–423 (2009)
23. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. **25**(11), 1251–1255 (2007)
24. Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome biology **6**(5), R46 (2005)
25. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids Res. **37**(suppl 2), W170–W173 (2009)
26. Salvadores, M., Alexander, P.A., Musen, M.A., Noy, N.F.: Bioportal as a dataset of linked biomedical ontologies and terminologies in RDF. Semantic Web **4**(3), 277–284 (2013)
27. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)

28. Dürst, M., Suignard, M.: Internationalized resource identifiers (IRIs). Technical report, RFC 3987 (2005)
29. Klyne, G., Carroll J.J.: Resource description framework (RDF): concepts and abstract syntax (2006)
30. Sporny, M.: JSON-LD and why I hate the semantic web. http://manu.sporny.org/2014/json-ld-origins-2/ (2014). Accessed 19 Jan 2015
31. Dolin, R.H., Alschuler, L., Boyer, S., Beebe, C., Behlen, F.M., Biron, P.V., Shabo Shvo, A.: HL7 clinical document architecture, release 2. J. Am. Med. Inform. Assoc. **13**(1), 30–39 (2006)
32. Operational data model. http://www.cdisc.org/odm (2015). Accessed 20 Jan 2015
33. Gadde, S., Aucoin, N., Grethe, J.S., Keator, D.B., Marcus, D.S., Pieper, S.: XCEDE: an extensible schema for biomedical data. Neuroinformatics **10**(1), 19–32 (2012)
34. Edwards, P., Mayernik, M.S., Batcheller, A., Bowker, G., Borgman, C.: Science friction: data, metadata, and collaboration. Soc. Stud. Sci. **41**, 667–690 (2011). doi:10.1177/0306312711413314
35. Neu, S.C., Crawford, K.L., Toga, A.W.: Practical management of heterogeneous neuroimaging metadata by global neuroimaging data repositories. Front. Neuroinform. **6**, 8 (2012)
36. Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L.: The extensible neuroimaging archive toolkit. Neuroinformatics **5**(1), 11–33 (2007)
37. XML schema. http://www.w3.org/XML/Schema (2014). Accessed 20 Jan 2015
38. Herrick, R., McKay, M., Olsen, T., Horton, W., Florida, M., Moore, C.J., Marcus, D.S.: Data dictionary services in XNAT and the human connectome project. Front. Neuroinform. **8**, 65 (2014)
39. Scott, A., Courtney, W., Wood, D., De la Garza, R., Lane, S., King, M., Wang, R., Roberts, J., Turner, J.A., Calhoun, J.D.: COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. Front. Neuroinform. **5**, 33 (2011)
40. COINS - central authentication system. https://www.coins.mrn.org/dx (2015). Accessed 05 Feb 2015
41. Austin, J., Jackson, T., Fletcher, M., Jessop, M., Liang, B., Weeks, M., Smith, L., Ingram, C., Watson, P.: CARMEN: code analysis, repository and modeling for e-neuroscience. Procedia Comput. Sci. **4**, 768–777 (2011)
42. Corradi, L., Arnulfo, G., Schenone, A., Porro, I., Fato, M.: XTENS - an extensible environment for neuroscience. Stud. Health Technol. Inform. **147**, 127 (2009)
43. Corradi, L., Porro, I., Schenone, A., Momeni, P., Ferrari, R., Nobili, F., Ferrara, M., Arnulfo, G., Fato, M.M.: A repository based on a dynamically extensible data model supporting multidisciplinary research in neuroscience. BMC Med. Inform. Decis. Mak. **12**(1), 115 (2012)
44. XSL Transformation. Version 1.0, W3C recommendation 16 November 1999
45. Krestyaninova, M., Zarins, A., Viksna, J., Kurbatova, N., Rucevskis, P., Neogi, S.G., Gostev, M., Perheentupa, T., Knuuttila, J., Barrett, A., et al.: A system for information management in biomedical studies SIMBioMS. Bioinform. **25**(20), 2768–2769 (2009)
46. Bauer, C., King, G.: Hibernate in Action. Manning, GreenWich (2005)
47. SIMS configuration guide. http://www.simbioms.org/wordpress/wp-content/uploads/2013/08/sims_configuration_guide_02.14.pdf (2013). Accessed 20 Jan 2015
48. Bauch, A., Adamczyk, I., Buczek, P., Elmer, F.-J., Enimanev, K., Glyzewski, P., Kohler, M., Pylak, T., Quandt, A., Ramakrishnan, C., et al.: openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinform. **12**(1), 468 (2011)
49. Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S., Kohane, I.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J. Am. Med. Inform. Assoc. **17**(2), 124–130 (2010)
50. Kimball, R., Ross, M.: The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley, New York (2011)
51. Segagni, D., Tibollo, V., Dagliati, A., Zambelli, A., Priori, S.G., Bellazzi, R.: An ICT infrastructure to integrate clinical and molecular data in oncology research. BMC Bioinform. **13**(Suppl 4), S5 (2012)

52. Natter, M.D., Quan, J., Ortiz, D.M., Bousvaros, A., Ilowite, N.T., Inman, C.J., Marsolo, K., McMurry, A.J., Sandborg, C.I., Schanberg, L.E., et al.: An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J. Am. Med. Inform. Assoc. **20**(1), 172–179 (2013)
53. Crockford, D.: JSON: The fat-free alternative to XML. Proc. of XML **2006** (2006)
54. Iacob, E.: The extended XPath language (EXPath) for querying concurrent markup hierarchies. http://dblab.csr.uky.edu/~eiaco0/docs/expath (2005)
55. Chen, R.S., Nadkarni, P., Marenco, L., Levin, F., Erdos, J., Miller, P.L.: Exploring performance issues for a clinical database organized using an entity-attribute-value representation. J. Am. Med. Inf. Assoc. **7**(5), 475–487 (2000)
56. Dinu, V., Nadkarni, P.: Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. Int. J. Med. Inform. **76**(11), 769–779 (2007)
57. JSON. http://json.org/ (2015). Accessed 23 Jan 2015

http://www.springer.com/978-3-319-31240-8

Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management

Izzo, M.

2016, XX, 104 p. 27 illus., 16 illus. in color., Hardcover
ISBN: 978-3-319-31240-8