Semantic Representation of Provenance and Contextual Information in Scientific Research

DISSERTATION

zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

Im Fach Bibliotheks- und Informationswissenschaft

eingereicht an der
Philosophische Fakultät I
Institut für Bibliotheks- und Informationswissenschaft
Humboldt-Universität zu Berlin

von
Dipl. M.Sc. Armand Brahaj

Die Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. habil. Dr. Sabine Kunst

Die Dekanin der Philosophischen Fakultät I
Prof. Dr. Gabriele Metzler

Gutachter/in: 1. Prof. Dr. Peter Schirmbacher
2. Prof. Dr. Vivien Petras
3. Prof. Dr. Detlev Doherr

Datum der Einreichung: 16.06.2016

Datum der Promotion: 19.09.2016

This document was created on: 14.11.2016

Abstract

Computational and information technology is one of the biggest advancement of the last century, a revolution that is influencing the way we approach social and technical problems in our day to day life. While these technologies have already influenced the research activity per sé, it is to be expected that these innovations will significantly influence the publishing and sharing of scientific results as well. So far, scientific publications have relied on limited result data attached inline in research paper publications. Establishments supporting research are pushing for concrete solutions that allow dissemination, share and reuse of research results. Reports such as "Riding the Wave - How Europe can gain from the rising tide of scientific data" of the High Level Expert Group on Scientific Data, European Commission (High Level Expert Group on Scientific Data, October 2010) presents a vision where the challenges of diverse data formats, people and communities are avoided due to the application of technical, semantic and social features of interoperability.

This research is an effort to address similar concerns from a technical perspective. Focus of this research is the exploration of a novel approach on supporting research data curation by developing a method and defining an automated data curation process where data can be easily annotated. As a contribution, this work offers a formal model (COSI) that allows integration of plentiful metadata that can be treated as logic concepts and not merely as literals. These concepts are defined in an ontology that allows among other actions, inference and reasoning operations. The second contribution of this work is associated to a pragmatic solution that facilitates annotation of metadata on the fly. This solution is referred as sheer curation and shows how data can be annotated (based on COSI) and published while investigations are executed. This research offers a creative solution that allows researchers to annotate and publish easily their scientific investigation data. This thesis offers a pragmatic a model, implementations and technologies that simplify the scientific data management and scientific publishing of research data.

Zusammenfassung

Die Computer- und Informationstechnologie ist eine der größten Errungenschaften des letzten Jahrhunderts -- eine Revolution, welche die Art und Weise beeinflusst, auf die wir im täglichen Leben auf technische und soziale Problemen reagieren. Obwohl diese Technologien bereits Forschungsaktivitäten an sich beeinflussen, so ist zu erwarten, dass sie auch einen Einfluss auf das Publizieren und Teilen von Forschungsergebnissen haben werden. Bisher wurden in wissenschaftlichen Publikationen nur in geringem Maße Daten beigefügt. Forschungförderungseinrichtungen drängen zu konkreten Lösungen zum Verbreiten, Teilen und Wiederverwenden von Forschungsergebnissen. Berichte wie "Riding the Wave - How Europe can gain from the rising tide of scientific data" der High Level Expert Group on Scientific Data der Europäischen Kommission zeichnen eine Vision, bei der die Herausforderungen einer Diversität an Datenformaten, Menschen und Gemeinschaften durch die Anwendung technischer, semantischer und sozialer Eigenschaften der Interoperabilität vermieden werden.

Diese Forschung adressiert derartige Herausforderungen aus einer technischer Perspektive. Fokus dieser Arbeit ist die Exploration eines neuartigen Ansatzes zur Unterstützung der Kuration (Sichtung und Korrektur) von Forschungsdaten mittels der Entwicklung einer Methodologie und mittels der Definition eines automatischen Datenkurationsprozesses in welchem Daten auf einfache Weise annotiert werden können. Ein Beitrag besteht in einem formalen Modell (COSI), welches die Integration großer Mengen an Metadaten erlaubt, welche als logische Konzepte behandelt werden können anstatt nur als Literale. Diese Konzepte werden in einer Ontologie definiert, welche, unter anderem, Inferenzen und Schlussfolgerungen ermöglicht. Der zweite Beitrag dieser Arbeit besteht in einer pragmatischen Lösung die es erlaubt, Metadaten on-the-fly zu annotieren. Diese Lösung wird als *Sheer Curation* bezeichnet und zeigt, wie Daten basierend auf COSI annotiert und publiziert werden können während Untersuchungen ablaufen. Die vorliegende Forschung bietet eine kreative Lösung, welche Wissenschaftlern Annotation und Veröffentlichung ihrer Investigationen ermöglicht. Diese Arbeit bietet einige pragmatische Modelle, Implementierungen und Technologien zur Vereinfachung sowohl des Managements als auch der Publikation wissenschaftlicher Daten.

Table of Contents

“Nullius in verba”
“Take no man‘s word for it”

Moto of the Royal Society[1]

[1] The Royal Society's motto 'Nullius in verba' is an expression of the determination of fellows to withstand the domination of authority and to verify all statements by an appeal to facts determined by experiment (The Royal Society, 2013).

Researchers should be motivated to create metadata
immediately and tool developers should add those
descriptors that can be created automatically.
"It is known that if metadata is not created immediately at resource
creation time the costs will increase rapidly and the quality decreases
requiring costly curation efforts.

e-IRG Data Management Task Force „Report on data
management" (e-IRG Data Management Task Force, 2009)

1. Introduction

Research data management has received a lot of attention in recent years due to its importance to research activity and correlations to the quality of research itself. Advancements in Information Technology have transformed the way scientific activity is performed shifting the data management focus toward digital data management. Almost all scientific disciplines are facing a data deluge nowadays. And yet, scholarly communications have relied on limited data results attached inline in research reports or peer-review publications. Increasing demands by researchers in pursuit of trust and reusability is pushing toward the practice of sharing the entire range of research datasets generated in the course of investigations. Provenance and contextual information regarding scientific investigations can allow reproduction of the process and offer valuable feedback to follow up by other researchers. To comply with the need of sharing research data, researchers are urged to produce quality metadata as part of their work.

Sharing the complete datasets created in the course of research is becoming an indicator of good research practice. It is not only the ability to reproduce investigations that makes these research datasets valuable. Interoperability features allow researchers across different disciplines to access and make use of the valuable information. This has been recognized as e-Science, a concept related to the global collaboration of researchers and citizens (Hey, et al., January 2003). *Data science*, another new paradigm in the research practice where knowledge is extracted by analysing research data (Gray, 2009) and where data is seen as a primary product is also highly dependent on the quality of the data annotation. Quality of the annotation and contextual information provided with the research data allow for future processing and examinations. Accessibility, interoperability and discovery all depend on the quality and extent of the provided metadata. The increasing magnitude of datasets produced every day in research activities, this data deluge, leads to challenges related to organisation and quality of research datasets.

In addressing the challenge of data deluge in research data management, a lot of attention has been dedicated to the digital curation practice. Efforts and initiatives focused on digital curation can be found in different communities at an international level. Initiatives and recommendations such as *Digital Curation Centre*[2] or the *Commission Recommendation of 17.7.2012 on access to and preservation of scientific information*[3] (European Commission, 17.7.2012) address the life cycle of research data including acquisition, curation, metadata creation, provenance, persistent identifiers, authorisation, authentication and data integrity. Many of the procedures involved in digital curation and data management are not new. Unsurprisingly, significant initiatives come from the library studies and have been the foundation of a community of digital library researchers (contributing at the same time to the birth of the library and information science). This is mainly because data curation activities have been seen as extensions of cataloguing and preservation procedures, such as consistency in naming, efficient tracking of versions, ensuring ethical aspects are honoured and defining appropriate storage characteristics. Different information models have also evolved to improve these activities. The *Dublin Core*[4] initiative for example, is widely used in data annotations not only in digital libraries but also in other resource management repositories. Other models such as the *ISO/IEC 11179 Metadata Registry (MDR)*[5] or the *Data Documentation Initiative*[6] are developed to create standards to organize results and data information generated by studies.

Most of the standardization initiatives are domain specific. Due to a close relation with digital library practices, we can find a number of standards of digital curation related to disciplines that fit in the group of humanities and social sciences. The *e-Information Reflection Group Report on Data Management* (e-IRG Data Management Task Force, 2009)

[2] Digital Curation Centre - http://www.dcc.ac.uk/about-us

[3] "Commission Recommendation of 17.7.2012 on access to and preservation of scientific information" - http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

[4] DCMI Home: Dublin Core: Metadata Initiative (DCMI) - http://dublincore.org/

[5] ISO/IEC 11179 Metadata Registry (MDR) standard - http://metadata-standards.org/11179/

[6] Data Documentation Initiative - http://www.ddialliance.org/

provides an exhaustive list of information modelling initiatives. The same report, discussing coverage for social and non-social sciences, states that: *"In the social sciences, data centres regularly employ the Data Documentation Initiative (DDI) standards. [but] No general model for the representation of scientific metadata exists"*. As we can see, the information models used in social sciences have not satisfied the curation activities in the natural and life sciences. This can be related to the diverse nature of the final datasets in these disciplines and the difficulties to annotate them in a way that allows uniform interaction with the data. Nonetheless, regardless of the discipline, researchers should be able to understand the *attributes*, determine on the *quality*, and be able to trace *provenance* of data in order to produce valid and reliable answers to scientific challenges. Questions involved in this process might be: *What devices were used to capture the data (How were the data obtained?)? What fundamentals were used in the investigation? How were these devices configured (or calibrated)? What conditions influenced the investigation? Who conducted the investigation?*[7] *etc.* Making all these information explicit helps addressing one of the most demanding challenges in data reuse; *trust*. Aligning entities across different vocabularies and improving the discovery capabilities is also a desired outcome, which is related to the quality of the information provided together with the research datasets.

As the reader can easily anticipate, this work is related to Research Data Management. Operations related to Research Data Management cover different aspects. Schirmbacher in (Schirmbacher, 2015) separates three dimensions when dealing with this field. First dimension is the *Scientific Politics*[8] dimension, dealing with responsibilities and what are the political, financial and human conditions at regional, national or international level for such data management. The *Organizational* dimension deals with structures that need to be created and how the responsibilities and division of labour and scientists should be arranged!? And last but not least the *Technical* dimension. The technical dimension deals with technical aspects of the infrastructure and underlying technologies ensuring that an appropriate service in accordance with all legal conditions is provided.

This thesis presents a novel solution to the challenge of generating rich metadata and provenance information during investigations in structural sciences, a subset of Life Sciences and Natural Sciences. Life Sciences, Natural Sciences and other reference to scientific disciplines mentioned in this work are referenced based on the classification of Deutsche Forschungsgemeinschaft (DFG) as published in *DFG Classification of Subject Area, Review*

[7] A set of requirements and an analysis of the answers to be addressed in this work are to be found in Chapter 4

[8] The original term is coined in German as "Wissenschaftspolitische Dimension", which is not to be confused with the Scientific politics a late 19th-century political theory based on the philosophy of Auguste Comte, a sort of conservative liberalism.

Board, Research Area and Scientific Discipline (DFG, 2008). Focus of this work is related to curation possibilities for the metadata and contextual information that address aspects of ***trust, reusability*** and ***discovery*** *in research data management*. Discussing the quality of metadata, the authors of the report *e-Information Reflection Group Report on Data Management* (e-IRG Data Management Task Force, 2009) that *there is increasing pressure on researchers to produce quality metadata descriptions and that the creation should be done best at the point of resource creation*. Although quality is difficult to be defined per se, the importance of metadata is related to the understanding and reuse of the data being described. The quality of metadata descriptions is related to (1) *the comprehensive spectrum of provenance and contextual information provided*; at the same time, quality of metadata is also related to (2) *the underlying technology and the flexibility that this technology allows for use and reuse of the metadata*.

This thesis dwells upon a solution to the challenge of producing qualitative metadata, at their point of creation. It does so by providing an evaluation of a number of modelling technologies, in attempt to find the most suitable modelling technology and by providing a generic formalization that allows coverage and presentation of the core concepts related to the provenance and contextual information in investigations. Producing abundant and comprehensive spectrum of metadata can be cumbersome, therefore a solution that allows utilization of the formalized model and the underlying technology to produce metadata at the point of resource creation is also presented. The model and the implementation presented in this work embrace the recommendation of *e-Information Reflection Group Report on Data Management* and focuses on sheer-curation, a form of digital curation that addresses the creation of metadata during the execution of an investigation. The formalized information model is designed to describe and contain data deriving from scientific investigations. This model includes technical and social information related to investigations.

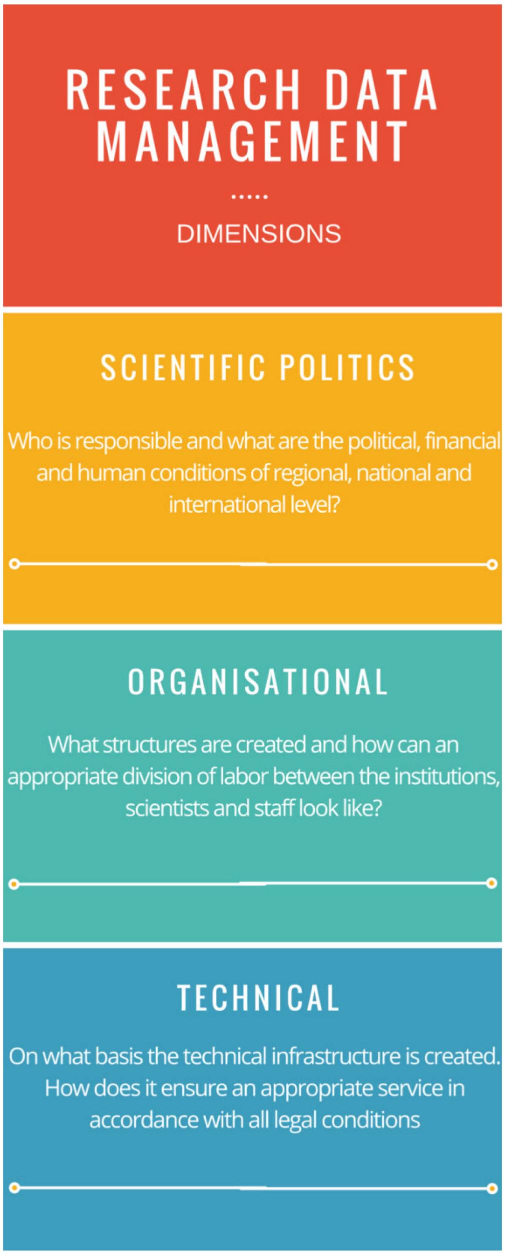

RESEARCH DATA MANAGEMENT
.....
DIMENSIONS
SCIENTIFIC POLITICS
Who is responsible and what are the political, financial and human conditions of regional, national and international level?
ORGANISATIONAL
What structures are created and how can an appropriate division of labor between the institutions, scientists and staff look like?
TECHNICAL
On what basis the technical infrastructure is created. How does it ensure an appropriate service in accordance with all legal conditions

Figure 1: Research Data Management Dimensions as defined by P. Schirmbacher **(Schirmbacher, 2015)**

It comprises information of the final environment where investigations take place and can easily be extended to be case specific. Key entities involved in this model are: *investigation (experiment, measurement, observation, trial), instruments, rigs, environment information, investigator, site, organization, study, programme* and more[9].

To address features of discovery and interoperability, the formalization is based on semantic technologies. Considering added values of semantic technologies in aligning concepts and entities, querying and reasoning capabilities, we discuss the potentials of such an approach across datasets of different disciplines and published in different repositories. To present a concrete implementation, eSciDoc infrastructure is used with the addition of a semantic index layer. The thesis presents a generic approach that is not coupled with a specific infrastructure and that can be used in combination to any repository system with minor alternations. The presented solution supports creation of the metadata and publication of these metadata in a repository allowing an automated form of data publishing and data sharing.

1.1 Significance of the Problem

The approach researchers follow to present new findings and making them acceptable in society has developed in its own methodology. The reception of new findings has always passed through rough filters before it was acknowledged as true. Acceptance of new knowledge is not a simple process for the intellectual beliefs; the process is a mini-revolution per sé. Presentation of new findings in science usually follows defined steps that will make it easier to break old intellectual beliefs in favour of the new ones. This process at times has shown to be fatal for the scientists if inadequately presented or left little impact just to be revived by other scientists to take all the credits due to a proper presentation and implementation.

The practice of presenting new findings has passed through many milestones in history. Probably the first we could trace is Aristotle and his empiricism (Gauch, 2003) (North, 2005). Avicenna rising on the shoulders of theological practices defined standards of how diseases, facts and patients-history should be recorded for later research (McGinnis, 2008). Roger Bacon shaped beautifully this methodology in the "scientific model" (Thorndike, 1914). Henry Oldenburg addressed the authenticity problem of research by producing the first "peer-reviewed" journal (Royal Society of London, 2015). Alexander von Humboldt showed the power of interconnecting scientific results across disciplines (von

[9] Entities and relations of the model are discussed in details in Chapter 4, *Scientific Investigations, Provenance and Contextual Information*

Humboldt, 2003) and the list can go on until more recent days with visionaries such as Vannevar Bush and his Memex machine (Bush, 1945), a source of inspiration for many modern knowledge organization systems or Tim Berners-Lee and the Semantic Web (Berners-Lee, et al., 2001).

Some argue that the biggest advancement of last century is the computational and information technology. It is to be expected that this innovation has significantly influenced the methodology of publishing scientific results as well. So far, scientific publications have relied on limited result data attached inline in research paper publications. While scholarly communication is an established practice aiming for research dissemination through journal articles, books or thesis, the publication and sharing practice of research datasets is still in the early phases. Attention to publication of research datasets is making it an integral part of the academic publishing due to its correlation with re-usability and credibility of the research process. The increasing attention to the publication of research datasets can be evidenced in the requirements of different journals and organisations. The *Nature* (Nature Publishing Group, 2014) journal for example, in their "Availability of data and materials" section demands that: "*Data sets must be made freely available to readers from the date of publication, and must be provided to editors and peer-reviewers at submission, for the purposes of evaluating the manuscript* (Nature Publishing Group, 2014). The same practice is being enforced by grant providers as well. National Institutes of Health in their "Sharing Research Data" policy state: *"The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. [...] investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing"* (National Institutes of Health, 2003). Similar directives are issued also by policy makers, such as the memo *"Increasing Access to the Results of Federally Funded Scientific Research"* (The White House, 2013) in the U.S.A or the "*Commission Recommendation of 17.7.2012 on access to and preservation of scientific information*" (European Commission, 17.7.2012) in the European Union.

The increasing attention to the process of sharing research results needs to be addressed by resourceful solutions from the information technology perspective as well. These solutions need to make it possible to disseminate individual data as independent publications and as standalone academic resources. At the same time these solutions need to be cost and resource effective, be complete and allow for further processing. Researchers should be attracted into providing their research results with the same incentives they have in academic publishing. These creative solutions should allow for the data sets to be findable, reusable, citable and allow attribution offering encouragements for researchers to make their research datasets publicly. New paradigms such as altmetrics[10] are already changing the way academic outreach is measured and this relies on the magnitude and quality of the shared data. Organizing research data on an atomic level opens new opportunities for research and data-science. Artificial Intelligence (AI) hype of the last 50 years did not deliver all the initial promises. Out of the diminishing AI, Semantic Web paradigm (also referred to as Web of Data) emerged. Semantic Web borrowed some solid concepts, such as reasoning and agent services from AI and together with an atomic representation of data and interconnection through WWW presents today a powerful consolidation of data that can be exploited in data research activities. With the right annotation of research data, and good workflows that assist publishing in public repositories, the semantic web presents an outlook where data create a large graph of knowledge containing valuable information for every research practice. We are living a mini-revolution in the way research is processed and quality and abundant information on the provenance of resources is vital to the process.

1.2 State of the Art

Formalization of knowledge acquisition has been a hot topic and a motivating force since ever. In the scope of this work, a set of predecessor models have been analysed. One of the earliest initiatives of using semantic technology to model knowledge acquisition has been the "Community is Knowledge! in *(KA)2*" (Richard, et al., 1998) initiative, referred to as (KA)2. (KA)2 was focused on the World Wide Web to formalize research teams, projects and scientific documents distributed as Web resources. As one of the earliest ontological models, it had a strong emphasis on the use of semantic technologies to formalize research activity.

Other projects have dealt more specifically with research data within specific scientific disciplines. In the field of biology for example, several initiatives have explored on modelling sets of discipline specific ontological models. *The Open Biological and Biomedical Ontologies*[11] for example is a collaborative effort involving many developers of science-based ontologies. The focus of the initiative is to establish principles for ontology development with the goal of creating a suite of interoperable reference ontologies in the biomedical domain. Similar ontologies follow a vertical knowledge management schema and are developed for a particular domain and specific situations.

[10] See Altmetrics: A manifesto - http://altmetrics.org/manifesto/ (Retrieved 2014)

[11] The Open Biological and Biomedical Ontologies official homepage - http://www.obofoundry.org/ The initiative will be referenced again in discussing BFO, one of its most prominent ontologies.

Some other initiatives aimed for a more generic ontological approach on the organization of experiments and research data. *EXPO* (Soldatova, et al., 2006) presents an ontology of scientific experiments aiming to formalise generic knowledge about scientific experimental design, methodology, and results representation. The focus of EXPO is solely focused on the technical entities of an experiment. EXPO can be considered as a strong solution for the technical representation of concepts contained in an experiment.

While the aforementioned models are all ontological models, there also have been attempts to formalize and present scientific activity through XML models. From these models, the *Core of Scientific Metadata Model (CSMD)* (Matthews, et al., 2010) aims to cover the general structure of scientific data holdings. CSMD was found very useful in the course of this research. It is designed to be a core system which is extensible and can be specialised to particular investigation independently of the scientific domain. As CSMD covers a general presentation of an investigation, it was ported to an ontology in the *Core of Scientific Metadata Ontology* (Brahaj, et al., 2012). CSMO was established as an orthodox representation of CSMD in an ontological model. Integrating CSMO to the workflow of an investigation in practice proved that the modelling was too restrictive in expressing the necessary investigation scenarios. As such, the technical communication on “BW-eLabs Report” (Razum, et al., 2012) presented a set of amendments incorporated in CSMO in the implementation of *BW-eLabs project*[12]. The model designed in the course of this thesis is a continuity of the work started in the *Core of Scientific Metadata Ontology* and *BW-eLabs* model. It is an ontological model that captures vivid metadata that can be used to improve findability, interoperability and alignment of scientific efforts across reposotories in different locations.

Another set of models encountered nowadays in the modelling of research activities are the provenance related models. One of the most prominent provenance vocabularies is the *Dublin Core* (DC) (The Dublin Core Metadata Initiative (DCMI) , 2012). DC Metadata Terms provide a set of basic metadata used to annotate resources. As metadata is not provenance, only a subset of the DC Terms can be used to track the provenance of a resource. Similar descriptions in DC answer questions such as - *Who created the resource*? *When was it changed?* etc. DC Metadata Terms are also ported to an ontological representation. Large EU research initiatives such as CIDOC-CRM (International Organization for Standardization, 2014) have focused on providing definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. *PREservation Metadata: Implementation Strategies* (Library of Congress, 2008), or better known by its

[12] The official webpage of the BW-eLabs is http://www.bw-elabs.org/. The BW-eLabs Report is referenced in the bibliography under index: (Razum, et al., 2012)

acronym PREMIS, is a data dictionary developed to support formalization of metadata related to long-term preservation. It defines a core set of semantic units that repositories should recognize in order to perform their preservation functions. It focuses on the provenance of the digital objects and not on the provenance of the descriptive metadata. *Provenir Ontology* (Sahoo, et al., 2009) is also an ontology targeting to model provenance. It is claimed to be more expressive in terms of the modelled concepts and well defined named relationships (Sahoo, et al., 2009). It is based on three base classes, *data, agent* and *process*. As we will see, these three classes are central to most of the provenance related vocabularies with slight terminological changes. In the case of *Open Provenance Model (OPM)* (Moreau, et al., 2011)**,** the basic classes are grouped under *process, artefact* and *agents*. OPM was the result of a set of activities known as the Provenance Challenge. The Provenance Challenge (Provenance Challenge, 2006) derived as a community effort from the *International Provenance and Annotation Workshop*[13] 2006. The aim of the challenge was to understand the different representations used for provenance. Its common aspects, and the reasons for its differences. OPM was superseded by the *Provenance Data Model* known as *PROV-DM* (Moreau, et al., 2013) and the Provenance Ontology, branded as PROV-O (Lebo, et al., 2013). PROV-DM is currently a W3C Recommendation and actually the reference model for many provenance implementations. The PROV-DM implementation page references more than 65 documented implementations in practice. The PROV data model is composed of core structures and other modular extended structures. Core structures are basic classes that form the essence of provenance information. Advanced uses of provenance can be achieved through integration of extended structures. Extended classes enhance and refine core structures with more expressive capabilities. The PROV data model has a segmental design and is structured in six components covering various facets of provenance. These components cover: *1) entities and activities together with the time at which they were created, used or ended; 2) derivations of entities; 3) agents bearing responsibility for entities that were generated and activities that happened; 4) bundles, a mechanism to support provenance of provenance;5) properties to link entities that refer to the same thing and 6) collections forming a logical structure for its members* (Moreau, et al., 2013).

PROV-DM improves over prior models as it provides guidance on the level of granularity that should be used in describing provenance of complex objects. The data published may contain many records composed of complex objects and provenance could be associated at any level of aggregation and granularity. Latest features of PROV-DM support as well activities and processes, a common standard for exposing and expressing provenance information that captures processes as well as the other content dimensions. PROV-DM and

[13] International Provenance and Annotation Workshop Series – The community page is www.ipaw.info. Accessed on March 2015.

PROV-O are also discussed in more details in the Chapter 4 focused on the engineering of an ontology as an information model for scientific investigations.

As we see, there are many proposed approaches and technology solutions that are relevant to provenance. Vocabularies and ontologies related to provenance cover fundamental concepts and activities of derivation nature. Nevertheless, these models are related to a broad generalization of provenance. In order to provide a relation between models describing experiments such as EXPO, or investigation activities such as CSMD, concrete extensions of provenance models are needed. EXPO or CSMD provide contextual information on the artefacts that are employed in an investigation, but they do not foresee the integration of provenance activities in their models. With regard to an ever increasing data oriented research activity, it is important to rely on modelling solutions that provide a comprehensive contextual and provenance information of the digital artefacts. This need is materialized in the research questions that are discussed in the next section.

1.3 Research Goal and Research Questions

The goal of this thesis is to present a set of models, implementations and technologies that simplify the scientific data management and scientific publishing of research data. We do so through the use of sheer curation and a specific information model based on semantic technologies. The presented methods are generic and not coupled with specific proprietary solutions. I also provide an assessment of the semantic annotation and representation of data focusing on advantages and disadvantages of the use of this technique as well.

The magnitude of data produced every year in research is increasing. By use of sheer curation and semantic annotation we can drastically improve the publication process and quantity and value of research data. The approach should not only improve the process but also open new possibilities in data-science. The research aim of this dissertation is focused on the thesis that: *Proper annotation of the result-data with semantic technologies, at their point of creation can produce abundant contextual information to allow reproduction of the investigation and provide a clear outlook of the process increasing trust in the research processes. In addition, the use of semantic technologies to annotate research-data improves visibility (findability) and usability.*

I substantiate the main research aim by the following set of research questions. Each research question highlights a different facet of the main claim, while focusing on providing a solution for the documentation and annotation of information related to a scientific investigation.

Research Question 1

How can we model the (finite) environment, including entities and relations that are part of an investigation process?

Such a model should allow for harvesting of provenance and contextual information containing information on entities such as institution, investigators, study, research and research results.

This research question is addressed in Chapter 4. This chapter is dedicated to an analysis for the evidence of core entities that should be used to describe provenance and contextual information from scientific investigations. Based on the results of the analysis, a semantic model is formalized. The benefits of the specific formalization are elaborated in the discussion on Chapter 3 on evaluation of an Information Modelling technology.

Research Question 2

In an ever-increasing data deluge and research activity based on digital environments: *How can we use the aforementioned formalization model to simplify the annotation process of research data? Is it possible to automate the process of data annotation and at what extent?*

I address this research question by presenting a semi-automation solution for the annotation of research data formalized in a semantic model. The presented solution is based on the practice of Sheer Curation[14] and allows for simple annotation of the whole metadata spectrum at the point of creation of the research data. This research question is addressed in Chapter 4, under Section 4.3, Sheer Curation.

Research Question 3

From Laboratory to Repository: *How to automate the publishing process of research data in data repositories and still comply with requirements of good scholarly communication practices? How can this formalization be used to improve the interconnectedness in research activities?*

Chapter 2 presents a set of requirements of that constitute a scholarly communication practices. While Research Question 2 addressed the automation of research data annotation process, in this research question the focus is on possible advancements on the publishing process of research data and metadata in (public available) repositories. The published assets should be citable, uniquely identified and their longevity history preserved, in the same way scholarly articles are treated. This research question discusses how the formalized model (see

[14] The term Sheer Curation is explained in Section 4.2.

Research Question 1), can be incorporated in a workflow that allows metadata annotation (see Research Question 2) up to the publishing of the results.

Advantages and disadvantages of the aforementioned techniques and their impact in scientific publishing are discussed as well in Chapter 6, Discussions. The same section provides a presentation of the research findings and possible portability of the formalized model to other scientific disciplines.

1.4 Research Contributions

The investigation of the outlined research questions has led to the following three main contributions of this dissertation, which also constitute the scientific accomplishment.

Contribution 1
Formalisation of an ontological model for the representation the research accomplished in the course of a scientific investigation

I provide a semantic model for the formal representation of the research and research data. The work covers core entities of a research investigation focusing on provenance and contextual information. The model is engineered in an ontology presented in details in Chapter 4. As a *core model*, the aforementioned model can be easily extended and aligned with other models.

Contribution 2
Implementation of a generic solution that allows a semi-automated process of data annotation and automated ingestion in data repository based on the formalized model.

This contribution is also related to the desired annotation and preservation of metadata during the execution of an investigation. It is possible only for investigations executed in the presence of a digital environment and it affects the volume and quality of data curated. At the same time, it is possible to facilitate data publication and data sharing of final research result sets. The contribution consists of a generic approach (also decoupled from any VRE or similar environments such as eSciDoc) on how a virtual research environment or a small-science investigation environment can make use of the sheer curation process generating quality semantic annotated research-data. This contribution is described in Chapter 4. Initial feedback by domain experts was received from the publication (Brahaj, et al., 2012). This publication presented a minimal and early iteration of the model at hand and targeted the community of digital library experts at the International Conference on Theory and Practice of Digital Libraries (TPDL). The vision of the work was presented at the International Conference on

Knowledge Management and Knowledge Technologies aiming to receive feedback from the Knowledge management community (see (Brahaj, 2012)).

Contribution 3

Assessment of the advantages and limitations of Semantic Technologies use in Research Data Management.

We discuss the benefits of semantic technologies in scientific data management. This contribution is related to the vision of a large knowledge graph that can be based on different existing repositories of different disciplines and institutions. Challenges and potentials of data alignment, data navigation, data protection are part of the discussion. Practical implementations show the potential of the technology. An evaluation and discussion on the current difficulties is also presented.

1.5 Publications

Below is a list of publications that are related[15] to the contributions of this thesis:

Brahaj, A., Razum, M., & Hoxha, J. (2013). Defining Digital Library. In Research and Advanced Technology for Digital Libraries (S. 23-28). Springe Berlin Heidelberg: Springer-Verlag Berlin Heidelberg.

Brahaj, A., Razum, M., & Schwichtenberg, F. (2012). Ontological Formalization of Scientific Experiments Based on Core Scientific Metadata Model. In Theory and Practice of Digital Libraries (Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings Ausg., S. 273-279). Springer Berlin Heidelberg: Springer-Verlag Berlin Heidelberg. (Brahaj, et al., 2012)

Brahaj, A. (2013, September). Capturing and Sharing Scientific Research Data. In “Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies” (p. 31). ACM.

A list of publications also published during this PhD research:

[15] The list of publications includes a number of papers published in the course of this thesis that are aligned with the topic of research. Feedback by the community is deemed highly important to assess the neccessity of this specific research work. For this reason, publication (Brahaj, 2012) presents a vision of how the solution should look like. Publication (Brahaj, et al., 2012) presents a limited (in the context of a specific project) formalization model, while publication (Brahaj, et al., 2013) deals with a cross referenced topic, that of digital libraries and repositories. In this research work, these publications are referenced in the same way as there are referenced other intellectual works by other authors; content of these publications is not used, although the intellectual work is to be considered a continuity.

Brahaj, A., Doherr, D., & Hoxha, J. (2011) Behavior-Based Information Seeking in Digital Libraries. In Knowledge Generation, Communication and Management: KGCM 2011, KGCM

Hoxha, J., & **Brahaj, A.** (2011, September). Open Government Data on the Web: A Semantic Approach. In Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on (pp. 107-113). IEEE.

Hoxha, J., **Brahaj, A.,** & Vrandečić, D. (2011, September). open. data. al: increasing the utilization of government data in Albania. In Proceedings of the 7th International Conference on Semantic Systems (pp. 237-240). ACM.

Brahaj, A. (2010) Virtual Research Environment in the Digital Library of Alexander Humboldt. IAF, Beiträge aus Forschung und Technik, 2010, S. 83-85, IAF- Offenburg

1.6 Thesis Structure

This document is organized in 6 chapters. Chapter **1** serves an introductory section, discussing the importance of the work, related work and dwelling into the main claim of the dissertation and a set of research questions. Chapter **2** is dedicated to the foundations and premises of this research work. A discussion and analysis on information modelling techniques is discussed in Chapter **3**, with an evaluation on the most appropriate modelling technique to support thesis' claim. Chapter **4** is dedicated to the design and development of an information modelling with focus the organization of data related to scientific investigations. Chapter **5** is dedicated to an evaluation of the ontology and presentation of practical implementations of the recommended solution. A summary of the results of the thesis and the impact is discussed in Chapter **6**.

Chapter 1

The first chapter, (this chapter) is dedicated to an introduction in the scope of this work. The chapter contains 6 sections discussing in the first 2 sections the importance of the work and the state of the art or similar research activities. The main claim and research questions of this work are stated in Section **1.3** followed by a description of the thesis contributions in Section **1.4**. The last two sections offer a list of author's publications and this guide to the reader.

Chapter 2

This chapter reviews fundamentals in scientific data management considered in this work. It is a bridge between the state of the art practice, the potentials to be exploited and a backbone for the research contributions.

Key concepts such as Metadata, Provenance and Contextual information are defined in the first section of the chapter. As the work is deemed to impact the scientific publication standard, the Section **2.2** offers a discussion on scientific publication in general and then narrowing our focus in natural science model of research. The discussion is focused on the scientific method and scientific publication practices.

Section **2.3** is focused on data intensive science as one of the paradigm of future scientific investigations. The section discusses the vision of data-intensive science and motivations for digital curation and quality data annotation. Research Data Management is discussed in a special section focusing on key concepts and activities such as provenance, contextual information, data quality, data licensing, ingesting etc. These activities are also addressed in the contribution of this work, so the section includes definitions and techniques used nowadays in research data management.

Under the vision of Linked Data, Section **2.4**, a detailed discussion is oriented toward scientific data interoperability and data sharing best practices. Semantic Web as a technology and its foundations are covered briefly to present to present the vision in the scope of this thesis. This last section covers information semantic ontologies, knowledge representation, graph searching, graph aligning representation of data in these technologies.

Chapter 3

This chapter is focused on the selection of an appropriate methodology to challenge and prove the thesis main claim. A set of requirements are documented based on interviews with researchers and laboratory assistants. Section **3.1** is a presentation of the requirements which define the thesis contribution. The requirements are a result of scientific data management activities in different stages such as modelling research results, live data retrieval from investigation with initial (semi-automatic) data annotation, data curation, ingest data management and data interoperability.

Section **3.2** is centred on the analysis of three data model technologies which can be used to address the requirements. It provides a discussion on modelling techniques and considerations reflected in selecting a model which can best used to address the research questions. Section **3.3** provides further information on Ontological Data Modelling technology, the preferred information modelling technology in addressing the research questions of this dissertation.

Chapter 4

This chapter is organised in 5 sections. Section **4.1** Model Conceptualization covers the analysis and the rationale for the developed data model. Section's **4.2** main discussion is focused on the formalisation of a comprehensive ontological data model describing the assets employed in the course of a scientific investigation. The process of generating research data, automatic annotation and curation solutions are discussed in Section **4.3** Sheer Curation. Two critical aspects of data reuse, the licencing and access control to datasets modelled with the developed ontology are discussed on Section **4.4** and **4.5**.

Chapter 5

Chapter **5** is dedicated to an evaluation of the developed data model (ontology). The first section of the chapter is dedicated to a description of the methodology used, while the rest of the sections cover in details specific aspects of the evaluation.

The objective of this chapter was to evaluate COSI to ensure that the developed ontology is of an acceptable quality. The evaluation is done by following the framework presented by (Vrandečić, 2010). This framework is based on the evaluation of 23 methods that are based on six aspects of an ontology. Practical use cases and implementations of the presented solution are discussed in Section 5.8. Results found by the evaluation of each of the methods are presented in a chart to provide a summary of the COSI evaluation.

Chapter 6

This chapter summarizes the contributions of this thesis. The chapter is organized in 5 sections with the first sections focussed on the Research Findings, their significance and limitations of the approach in sections **6.1**, **6.2** and **6.3** respectively. The chapter presents s mapping of the developed ontology with ontologies used in other scientific disciplines. The chapter closes with a discussion of implications of the work in the foreseen future.

Scientific knowledge is communicated through scientific literature.
Knowledge is ultimately derived from data."

Klump et al (Klump, et al., 2006)

2. Foundations

A fundamental characteristic of our times is information technology empowered by the use of digital computing. The digital revolution and the widespread presence of digital computing devices have impacted science, economy and social aspects of our daily lives. In the realm of science, information technology has created new opportunities due to advancements on computing power and information organization. This has influenced the way we perform research, run investigations and deduct new knowledge. Digital management and scientific knowledge organization have inspired many models of scientific collaborations and knowledge representation in the last decades. Even the World Wide Web was born in a research facility and was motivated as a facilitator of scientific information organization and information exchange.

Since its birth in the 90s, the Web has become an essential mean of communication and an indispensable tool for businesses, social interaction and scientific communications. As the inventor of World Wide Web, Tim Berners-Lee points out, *the Web was envisioned as an information space, with the goal that it should be useful not only for human-human communication, but also for devices that would be able to participate and help* (Berners-Lee, et al., 2001). The observation of Berners-Lee and his work on advocating a Web of Data, point to further developments needed with regard to a scientific usage of the Web. Tim Berners-Lee and many others use the terms *Semantic Web* and *Web of Data* interchangeably.

The reference *Web of Data* is mainly used to denote the differences to the Web of Documents. Traditionally, the Web has been oriented toward the references and hyperlinks between documents. These documents would be represented by HTML, Media files or other dynamically generated (document) representations and less about the relation between data. With regard to scientific communication, the exchange of data is a vital activity. It is important not only for the trust in scientific claims, but also a valuable resource that can provide interoperability across disciplines. The volume of data generated in scientific investigations is progressively increasing in the last decades (Hey, et al., January 2003). Nowadays scientific instruments and computer simulations are the new standard for scientific methods to analyse and run experiments in many disciplines. The use of computers in science has changed the way scientists handle studies, retrieve and publish information related to their studies. Computer simulations are creating vast amount of data that require new scientific methods to analyse and organize them. The volume of data produced by modern research is increasing continuously and so is the number of experiments run in laboratories (Gray, et al., 2005) (Hey, et al., January 2003). Most of these data are exchanged through the Web, although they do not exploit fully the vision of Web of Data.

With the increase in volume of produced data and experiments, new challenges and opportunities arise in the process of collecting, organizing, interpreting and sharing these data. In order to exploit and reuse the data generated from the digitally run experiments, computer simulations, and digital surveys, scientists need to rely on new forms of information discovery, information alignment and data interoperability. New methodologies that support data mining and data visualization need to be envisioned. These methods should make it easy to understand data, derive knowledge and reprocess them in new investigations. In order to analyse, evaluate and reuse these vast digital data, proper metadata need to be provided. These metadata need to incorporate important information such as provenance, context or licencing. The presence of the appropriate metadata increases the value of digital research data. Well-structured and qualitative metadata annotation practices influence the ability to locate and mine research data from different disciplines and repositories. Requirements for data interoperability and collaborative science applications are also pushing the focus of data modelling techniques toward the concept of interlinked data. Needless to say that the revolution in digital technology will also shift the way we work, organize and run analytical operations on data and metadata. These changes need to be reflected in the daily research routines and impact research environments as well as the digital repositories.

This chapter is dedicated to a presentation of fundamental concepts and terminology clarification referred in the scope of this work. First section of the chapter provides a basic presentation of the discipline of *Research Data Management* with emphasis on data curation, and clarification on the used terminology for metadata. The importance of integrating metadata curation capabilities in research workflows is described in the following section related to *Scholarly Communication*. This section covers a brief presentation of the scholarly communication, pointing to the changes that this process and the scientific method will undergo with the increasing necessity of digital preservation of metadata and other information on research process. As a preamble to questions about interoperability and data science discussed in the *Introduction*, the chapter continues with a presentation of *Data Intensive Science* and *Linked Data*.

2.1 Research Data Management and Metadata

In the scope of this work, the term *digital data* and *data* are used interchangeably to refer to *any identifiable information represented as a digital asset*. Data in digital representation are crucial nowadays due to the increasing dependency on information technologies. Whenever discussing on data related to a research process, the term *digital research data* is used. Kindling & Schirmbacher (Kindling, et al., 2013) in defining the term explain that *"[U]nder digital research data we understand all data in digital form, which arise during or are result of a research process"*[16]. A similar definition is also stated in the volume *Research Data Management* (Higgins, 2012) where Higgins describes the term as: „*Data produced throughout the research lifecycle* [which] *includes any information in binary digital form that is created, stored, accessed and rendered with the use of computer technology*". With these explanations we can relate the term *data* to any discrete *digital representations*. Such representation might be a set of raw bits or well defined data organizations such as text files, images, word processor or other proprietary files. Combinations of digital assets or collections will be considered digital data as long as they can be addressed in a unique identifiable way. Examples are folders or URI endpoints pointing to a dataset container or a service interface.

The organization and preservation of data are subject to the field of research data management. The term *Research Data Management* is used to describe the activity that supports the allocation, the generation, processing and enrichment, archiving and publishing of digital research data itself or by a corresponding classical text production (Kindling, et al., 2013). The practice has been related to the set of policies, formal procedures, practices and implementations needed to manage the information lifecycle of an *enterprise*. As it is expected, research data management is concerned with activities and policies related to data lifecycle management, data policies, sustainability, administration roles, responsibilities and of course infrastructures. In the description above, the term *enterprise* is used intentionally to

[16] The original citation is in German: „Unter digitalen Forschungsdaten verstehen wir alle digital vorliegenden Daten, die während des Forschungsprozesses entstehen oder ihr Ergebnis sind." (Kindling, et al., 2013)

allow generalization of the concept. The enterprise of application might be a research investigation, a project, an institution, a government branch or a business structure. The nature and attributes of the *enterprise* defines further technical aspects of data management. Despite the internal organization and layers of the research data management, the central point of interest for research data management is the *data* and the way data and its alternations are preserved. These alternations on the data and other contextual information are contained in *metadata*. Metadata is a commonly known term pointing to data files containing information about other data. As metadata is supposed to provide information based on specific contexts, the metadata are broad and diverse. The value of metadata is also assorted. In many occasions this value is related to metadata use as means for discovering data objects, in other occasions the value of metadata is exploited by the precious provenance evidence. In such case the metadata help address questions related to history, modifications or other information such as experimental parameters, creation conditions etc.

Metadata

In the simplest description, *metadata* is defined as *data about data*. Although this is an accurate explanation for the term, an ever increasing attention to data science and data management practice has pushed toward further classifications of metadata. When researchers and data scientists discuss metadata, they expect some additional clarification of the specific metadata required in a specific context of use. This is probably because metadata is a generalized term and seems to have been a general concept since ever. With roots in old Greek, (or an adjacent culture), the term *meta* denotes the use of a concept in participation *with* some other concept or context[17]. As traced in Online Etymology Dictionary, *meta* has been interpreted as *beyond*, *after* or *behind* (Douglas , 2001). Due to the widely spread of the term *metaphysics*, the *meta* prefix has been sometime wrongly interpreted as *higher than, transcending, overarching, dealing with the most fundamental matters of.* Important is that the etymology of the word *meta (=with)*, indicates use of a concept in a specific context. In the case of *metadata*, the word points to information that will allow the interpretation of data in a specific context.

In scope of digital data management, we will use the term metadata to refer to structured information used to describe an actual stand of digital data. This information may be used to present evidences related to preservation activities, provenance and findability, but depending on specific use cases, metadata may include additional context based information

[17] A very accurate explanation of the term is found in the *Henry George Liddell, Robert Scott, A Greek-English Lexicon, on Perseus Digital Library* (Liddell, et al., 1920). Beyond the Greek description, the *me* – is accepted as a proto-indoeuropean term equivalent to the old English *mið*, or German *mit* (Douglas , 2001). It is still used in that context in Albanian *me* and Greek *με* and is crucial to the unterstanding of the term.

related to the data examined. In attempt to explain the domain of use for metadata, different categorisation efforts have been attempted.

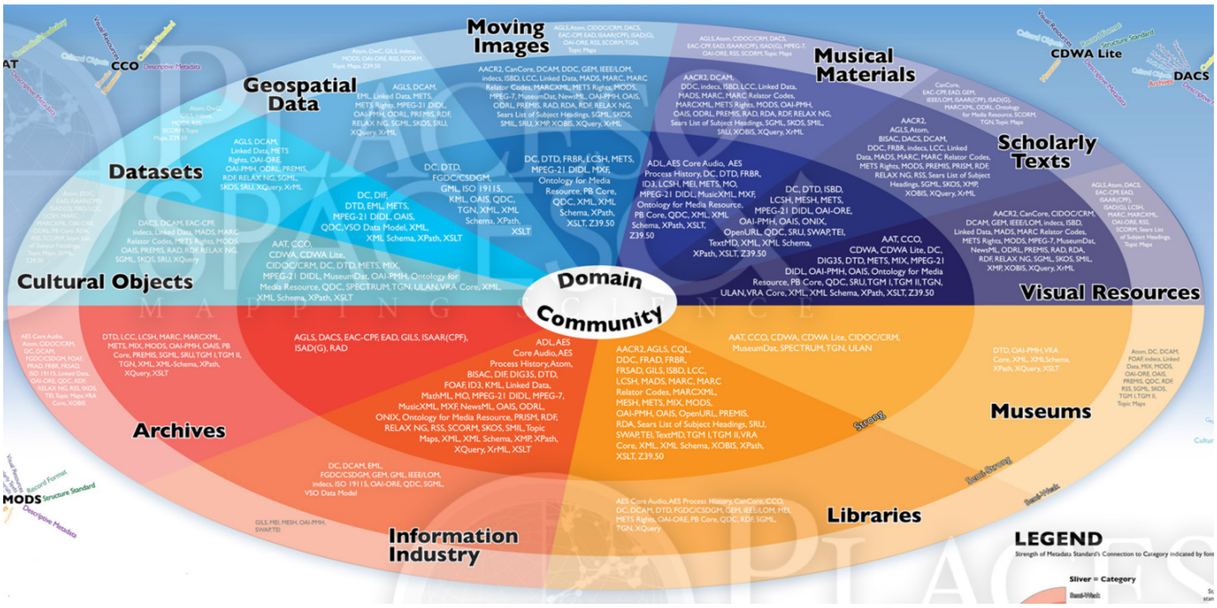


Figure 2: Visualization of the Metadata Universe **(Riley, et al., 2010)**

Figure 2 presents one of the graphs generated for *Visualization of Metadata Universe* (Riley, et al., 2010), an initiative originating at Indiana University. Visualization of Metadata Universe provides groupings of metadata in 25 different categories. These categories are defined based on four groups: *Domain, Community*, *Function and Purpose*[18]. The initiative is based on more than 100 different metadata standards which are mapped to these four groups.

Examining the metadata categorization from a functional perspective, one may notice that the Metadata Universe offers a grouping that includes *Technical Metadata, Structural Metadata, Rights Metadata, Preservation Metadata* and *Descriptive Metadata*. With minor alternations this metadata organization is also found to be promoted by other institutions such as the Harvard Library (Harvard Library, 2015). The categorization is an attempt to fragment the different metadata standards based on their functionality, from discovery, to managing access, to provision of information that enables the preservation and reusability. To better understand this categorization, below is a presentation of the categories. The most prominent metadata standards are pointed for each group.

Technical Metadata offer information on how a digital object was created, its format, specific technical characteristics and other technicalities on how to preserve reusable

[18] Figure 2 contains only the categorization based on Domain and Community. The Function and Purpose grouping can be found in the Metadata Universe visualization (Riley, et al., 2010).

artefacts. Metadata grouped under this category aim to provide the necessary information needed to manage digital objects over time. Some examples of prominent technical metadata are: TextMD[19] used for text files, AES Core Audio[20] used for audio files, MIX[21] for still images etc.

Structural Metadata is a broad category which includes metadata aiming to facilitate location and presentation of digital objects. Similar metadata may provide information about the internal structure of resources, describe relationship among materials and bind digital assets to a central object. The most prominent metadata standard for this category is METS[22], which can be used to aggregate related metadata. Other common structural metadata schemes MPEG-21: OAI-ORE[23] etc.

Rights Metadata offer information on copyright and protection of digital resources. Sometime this set of metadata is organized under the association **Administrative Metadata** including additional access control and quality related schemas. Although the Administrative Metadata categorization seems more meaningful than the Rights Metadata, in the scope of this research no coherent and active metadata standards where found to represent this category. With regard to the Rights Metadata, there are many rights expression languages (RELs) and other rights metadata standards that are exclusively part of this categorization. While the copyright is a very important aspect when dealing with research data, more has to be done with respect to the metadata standards used in practice. Some of the rights metadata are: METSRights[24], copyrightMD[25] etc.

[19] TextMD is a schema used to describe technical characteristics of text, such as encoding, character set, language, script and markup language - http://www.loc.gov/standards/textMD/textMD.xsd

[20] AES Core Audio, or AES60-2011: AES standard for audio metadata , defines an XML schema for the technical characteristics of an audio object (analog or digital). - http://www.aes.org/publications/standards/search.cfm?docID=85

[21] Metadata for Images in XML Standard (MIX) is an XML schema for recording and exchanging still images - http://www.loc.gov/standards/mix/

[22] Metadata Encoding and Transmission Standard (METS) is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library - http://www.loc.gov/standards/mets/

[23] Open Archives Initiative - Object Reuse and Exchange is defines standards for the description and exchange of aggregations of Web resources - http://www.openarchives.org/ore/1.0/toc

[24] METSRights helps documenting minimal administrative metadata about the intellectual rights associated with a digital object - http://cosimo.stanford.edu/sdr/metsrights.xsd

[25] CopyrightMD is a proof of concept initiative that has identified key data elements for expressing copyright metadata - http://www.cdlib.org/groups/rmg/

Preservation Metadata is a category focused on content and organization agnostic metadata. While different categories such as Technical Metadata, Structural Metadata and Descriptive metadata contain information for the preservation activity, this category is focused on standards and essentials for the preservation of information for the long-term. Well known metadata standard for this category are PREMIS[26] and the Open Archival Information System[27] (OAIS).

Descriptive Metadata describe and identify information resources. The aim is to provide information on intellectual content of a digital artefact. This category of metadata is composed of some of the most standardized and well understood metadata. It has a long tradition due to the primary focus of catalogues and traditional libraries. The metadata under this group differ based on the needs of specific communities. As declared, these metadata types are important for the resource discovery and they may support various user tasks. Some of the descriptive metadata are MARC 21[28], MODS[29], Dublin Core (The Dublin Core Metadata Initiative (DCMI) , 2012), Encoded Archival Description (EAD) [30] and many more.

Many of the metadata standards included in the analysis of the Metadata Universe are found in more than one category. It is clear that the metadata standards are not exclusive in each category. For example, the *AES Process History* can be mapped to the categories *Technical Metadata* and *Preservation Metadata*. In the same categories are found *TextMD* and *MIX*. *OAI-ORE* is found in *Descriptive* and *Structural Metadata* categories. In the same categories is also found the *MusicXML* schema. Observing the relation of metadata standards across different categories, it is to expect that higher generalization classification is also meaningful. The research department at the Cornell University Library (Cornell University Library/Research Department, 2003) presents a simpler categorisation of metadata in

1) *Descriptive,*
2) *Structural*

[26] The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability - http://www.loc.gov/standards/premis/

[27] OAIS provides a reference model for an archival system designed to maintain access to digital resources and preserve them over time - http://public.ccsds.org/publications/archive/650x0b1.pdf

[28] MARC21 is the primary library standard for the representation and communication of bibliographic and related information in machine-readable form - http://www.loc.gov/marc/

[29] MODS was designed both to carry selected information transferred from MARC21 records and to support the creation of original resource description records - http://www.loc.gov/standards/mods/

[30] Encoded Archival Description (EAD) is a mark-up language for archival finding aids, that is, detailed descriptions of collections that contain a wide variety of materials, including letters, diaries, photographs, drawings, printed material, and objects - http://www.loc.gov/ead/

3) *Administrative Metadata*

This new categorization does include all the metadata standards presented in the Visualization of Metadata Universe as it is a higher level of generalization. As the focus of this research is tightly connected provenance, trust and reusability, the upcoming section is focused on consolidating the terminology on the metadata groups of interest in this work.

2.1.1 Provenance and Contextual Information

Metadata are very often related to provenance information. The connection of provenance and metadata is so strong that often the two are equated (Gil, et al., 2010). Provenance provides the fundamental information to track the history and assess the legitimacy of primary data. Information provided as part of provenance enables trust and allows reproducibility of data and results. These activities are of critical importance, especially in scientific data management practice. Although provenance is connected to metadata, provenance metadata are a subset of metadata that provide specific contextual information. Provenance and contextual metadata gathered during the execution of investigations are central to the topic of this research. In this section we discuss the use of the terms *provenance* and *contextual information* within the broader concept of metadata.

As presented in the previous section, categorization of metadata based on functional perspective differs among researchers and institutions. In some classifications, there is an in-depth categorization. In such case, special categories are defined and metadata standards can be found in more than one category. An alternative categorization was the one advocated by the Cornell University Library (Cornell University Library/Research Department, 2003) where the focus is on a high level abstraction of the metadata categories.

In the scope of this thesis, the terms *provenance* and *contextual information* are used to generalize two groups of metadata standards. Provenance will be used to present information on the derivation history of a digital artefact starting from the creation moment. Such information includes logs on records of the entities and processes involved in producing or delivering the digital artefact. It also contains information on predecessor objects and transformations that lead to the actual state of the data. Depending on the nature of the investigation, the processes and entities invoked in the creation of data may be simple ones, but they might be workflows and agents of a higher complexity. The influence of these factors on the history of the digital artefacts is related mainly to technical aspects. Therefore, in the scope of this work, *provenance metadata* is used to denote information provided primary as metadata pertaining to technical and inner structural information of digital artefacts. These metadata contain information on how a digital object was created, its format, specific technical characteristics and any information that presents knowledge on how to preserve reusable data objects. Additionally, these metadata may include information on location and representation possibilities of the objects. Digital object may be atomic or composites. In the case of composite objects, provenance metadata contains information that describes each component of the composite data object.

Table 1: Metadata grouping for Provenance and Contextual Metadata

Provenance Metadata	Contextual Metadata
Structural Metadata	Description Metadata
Technical Metadata	Administrative Metadata
Preservation Metadata	Rights Metadata

With these observations, the provenance metadata presented in Table 1, can be used as a categorization that includes *structural, technical* and *preservation metadata* as we have seen in the previous section. Metadata standards that adhere to this grouping are TextMD, AES Core Audio, MIX providing information on technical aspects of the data; METS, MPEG-21, OAI-ORE provide information on the internal structure of data resources and relationship of the data composites, while PREMIS, Open Archival Information System (OAIS) and others contain information on the essentials for the preservation of information and asset management.

Beside the provenance metadata, we can relate to another set of metadata standards that are grouped under the term *contextual metadata*. Another reason to differentiate between these groups is to make honour to the linguistic terminology used. *Provenance*, deriving from Latin *provenire* refers to *come forth*, or *point of origin*[31]. *Provenance* is used to point to any information on the history and origin of a data file with focus on understanding its current state of being. Any information on the intellectual presence and use of the data will be addressed by another set of questions that are related to the context or a description of the data.

The categorisation of contextual metadata comprises metadata that describe primary-data's intellectual content. Such information relates the data to a creator, contributes to the understanding of data though textual descriptions, points to the right user audience, provides relation of content to events or disciplines, assist to classification of the data in logically created structures (or organisations such as hierarchies) and so on. Probably the most known metadata standards of this group derive from the Dublin Core Initiatives (The Dublin Core Metadata Initiative (DCMI) , 2012), although DC is a very extensive standard that covers different aspects. Within the category of contextual metadata, we find metadata schema that

[31] *Provenance* might also derive from middle ages French *provenant* but yielding the same result as the lating provenance argumentation.

provide the necessary information on data identification and findability, such as EAD or other metadata standards that provide categorisation of the data or works such as VRA Core[32]. Information on intellectual property rights or other licencing are also part of the contextual information.

The terms *provenance* and *contextual metadata* will be used intensively in the upcoming sections of this thesis. As we will see, the support of information technology in running investigations nowadays has influenced the way investigations are executed and metadata are generated. Metadata contain information on the *investigation workflow*, but also derive from *data processing* and *data curation* phases. Considering the advancements in information technology, all these phases are also found in environments supported by software implementations. Generation of metadata in these phases, whenever supported by a digital environment, can be automated. A practical implementation showing how these metadata can be harvested in a small science environment is discussed in Chapter 4. The metadata generated in a digital environment can easily contain the necessary provenance information such as devices used, their configurations and parameters. When combined in the flow of a Virtual Research Environment, the data curation process can produce on-the-fly metadata related to the creator of the data, the group of research, the institution, the study and more.

A great deal of intellectual information can be manually added to these data in the course of the curation process. Part of the information contained in the metadata automatically generated in the course of investigation might be similar, e.g. *The name of the laboratory where the data are being generated*. This may seem redundant, but such information is very important in understanding the rationale behind the data and it is up to the modelling technique chosen to preserve these metadata to impact further reuse and exploitation of the data. This thesis exploits the use of semantic technologies in modelling metadata and presents a novel approach of metadata generation and annotation in the course of investigations. There is a strong emphasis on metadata generation in the research environment. The need for abundance, clarity and quality of information provided through metadata is directly related to the scholarly communication and as we will see in the next section, it is related to the scientific method as well.

[32] VRA Core is a data standard for the description of works of visual culture as well as the images that document them http://core.vraweb.org/

2.2 Scholarly Communication and Metadata

Since the 17th century, scholars have relied on printed forms[33] of publications to document and disseminate the results of their work. The first acknowledged medium for scholarly communication was the *"Philosophical Transactions of the Royal Society"* (Royal Society of London, 2015), a journal Henry Oldenburg founded while he was working for the *Royal Society*, a UK based Academy of Science[34]. Oldenburg invited scientists to submit articles for publication and organized a committee of academics who would judge on the genuineness of articles submitted. *Philosophical Transactions* was quickly accepted by a community of scientists including Newton, Faraday, Darwin and many others who had often avoided declaring their discoveries for fear that someone else would claim priority (National Research Council, 2009).

In order to allow the experts of the *Philosophical Transactions* to judge on the scientific value of the work, the scientists included abundant information describing their research activity. These data included descriptions of the scientific investigations, description of the research process and description of all the entities used and produced in the course of investigations. A quick glimpse on the archives of the "*Philosophical Transactions of the Royal Society*" will show that publications include detailed information on the scientific investigations. For example, analysing an article by Ramsay and Young "*A Study of the Thermal Properties of Methyl-Alcohol*" (Ramsay, et al., 1886) one could easily see how the article includes information on the process including descriptions of the instruments, information on how one can reproduce the experiment and verify the outcome (also documented in the publication). The attention toward evidences on the scientific results was also sculpted by the Royal Society's motto: *"Nullius in verba*", translated *"Take no man's word for it"* (Royal Society of London, 2015).

The practice defined by Oldenburg in assuring quality and originality in scientific communications is not much different to the peer-review system as we know it today. The practice involved consultations with a group of experts who would decide and approve articles and findings appropriate to be published in an academic medium. Our modern

[33] Fully aware of the fact that in the course of history, we have relied on different methods of knowledge documentation, in the scope of this section I discuss the standard methods of scholarly communication based on the Johannes Guttenberg revolutionary printing press device. More than seven centuries later, printing forms of scholarly communication still remains a standard, although it is foreseen that this standard shifts rapidly toward digital communication. A number of standard methodologies for conserving knowledge in digital formats are already established, but discussing them is not the focus of this section.

[34] Philosophical Transactions was initially a private venture of Oldenburg, but it was soon incorporated as an official press of the Royal Society.

scientific publication system follows the same practice in accepting value and quality of scientific work in scholarly communications.

Documentation of the scientific research in journals and other means of scholarly communications has become an integral part of the researcher's activity. The practice is so fundamental to the scientific activity, that some scholars have included it as part of the *scientific method*, a term used to denote a group of techniques used in scientific investigations, acquisition of new knowledge, or correction and integration of previous knowledge. Crawford and Stucki (Crawford, et al., 1990) in explaining the linear actions that constitutes the modern understanding of scientific method provide the following steps:

- Define the question, Define the idea
- Search the literature, formulate statement
- Form hypothesis
- Develop methodology
- Develop proposal
- Test hypothesis
- Perform experiment and collect data
- Analyse and interpret data
- Draw Conclusions
- **Publish results**[35]

Descriptions of each of the linear steps of the scientific method are usually reflected in the same order in academic articles, or technical papers. In such a typical research publication, a hypothesis is described complemented by prior research information focusing on the state of the art of the research in the subject. A detailed description of the investigation process or the description of the methodology used is presented followed by the data generated. The research process might be a set of deductions, pure observations or the description of an investigation process. The result of the investigation might be explicitly related to the data generated or to a synthesis of the data generated. The result of the work is claimed in the conclusion of the publication.

Description of this research process is crucial for the acceptance of the scientific work by reviewers and the broad community. Scholars and reviewers of the publication need to acquire the confidence that the process description is realistic and/or that it contains all the necessary

[35] Not all the documentations of the scientific method include the „publication of results“ as the last step. Aligned with the (Crawford, et al., 1990) this thesis focuses on the importance of publishing the results of each investigation, even in case of negative data, or data that did not satisfy the hypothesis due to the value contained in failure as well as in success investigations. The reusability of data should allow other researchers to avoid ending in the same negative data. One of the contributions of this thesis indicates how the publication of the results can be easily be incorporated in the flow of scientific investigations run in digital environments.

information that allows them to test it as realistic. In other words, they need to believe that the process can be reproduced.

As we saw, the *Philosophical Transactions* established a standard of publishing scientific findings. The articles published in this journal included the abundant information and "metadata" to prove the original claim. The term *metadata* did not exist at the time. In fact, it was coined many centuries later. For the first time, the term metadata is documented to have been used in 1973 (Linux Information Project, 2006) by Jack E. Myers. Nevertheless, it is obvious that scholars have been providing what we nowadays refer as *metadata* before the term was properly coined. Probably a necessity to introduce this new term was the rapid surge in the volume of data and descriptive data generated in digital environments. (The term metadata started to have a universal use in the 1980s about the same time information technology impacted the increasing rate of data produced. Metadata is in fact a trademark which was first used 23-09-1981 by Metadata Corporation. The trademark has been renewed in 1997 and is still a valid trademark, despite the broad use in practice (United States Patent and Trademark Office, 1998). Due to the widespread of the term, it might be the case that it has entered the public domain as a general term). While the volume of data generated in investigations, simulations and other scientific activity backed by digital infrastructure grew exponentially, scholarly communication conserved the traditional format. The information that would guarantee the trust of the experiments was still provided inline in the published articles. As one might imagine, complying with the space limitation of publication standards, it was done in restricted form. To overcome these problems, many projects and initiatives have focused on publication of the metadata corpora in public repositories[36]. Publishing data and the metadata corpora, should be done in such way that these data are reusable and contain information that describe and allow interconnection of the corpora to additional resources, collection of resources, devices, processes and people.

The data management scenarios described in this work present the curation and the management of scientific data collected in the course of a scientific investigation. The activity is described as it would be represented in a Virtual Research Environment, where all the linear steps of a scientific method can be followed and documented. The annotation technique used in the process should allow for reproducibility, interoperability and allow for exploitation of data science activities. A promise paradigm fur such interactions is the Linked Data and its underlying semantic technologies.

[36] Examples include public repositories specially dedicated to the research community such as Figshare (http://figshare.com/) although other repositories such as Github (http://github.com) have been used for the task. These are of course solution beyond the hundreds of institutional repositories. Other initiates such as Projekt Radar are also targeting research institutions: https://www.radar-projekt.org/

2.3 Data-Intensive Science

Nowadays, the increasing utilization of the computing power in many science disciplines has change the way investigations are run. Most of the data in scientific research processes are generated through simulations, measurements, experiments, observations and other forms of investigation. Instruments in laboratories produce vast amounts of born-digital data, even in the so-called 'small sciences'. Jim Gray places the scientific research nowaday in the period of "Data-Intensive Science" (Gray, 2009), an era where scientists are overwhelmed with data and the research potential lies in the information technology that supports data mining, data analysis, data visualization and exploration. The number and nature of data sets produced and processed in scientific research nowadays provides value which is beyond the modest presentation in scientific journals.

The problem of the magnitude of data produce is described in different articles such *The Data Deluge: an e-Science Perspective* (Hey, et al., January 2003) where the authors argue that there are immediate needs for new methods and organization of research data. This leads to challenges in the process of collecting, organizing, interpreting, publishing and sharing the data. Special focus is put on sharing of data as this is becoming a requirement from both donors and expert's community. The value of sharing the research data has a lot of benefits for science, might this be merely cost related, due to possible data reuse or what is more important, knowledge related and supportive for multidisciplinary research. The data-sharing practice is being enforced by donors and by scholarly communication media that require researchers to share their data with other investigators, either by hosting the data in public repositories, or making them available upon request (Savage, et al., 2009).

There is also increasing consideration toward the publication of the negative data, or the data that did not satisfy initial hypotheses of the investigations. Articles such as *Why publish your negative results!?* (Sprott, et al., 2012) show that these data are considered valuable for other or future investigations. The proficiency of research data-sharing, depends on the simplicity of solutions provided to researchers in the process of capturing and preserving investigation results. It is the organization of the digital environment where researchers operate, that can influence the quality and quantity of data stored. Information technology experts need to support the creation of tools that allow researchers to annotate, store and make their data publicly available with as little effort as possible and without altering their investigation environment. While recent studies show that even though everyday life moved into the digital age and almost everything is shared in the Web 2.0, it appears that sharing research data is not yet a common thing to do (Dallmeier-Tiessen, 2011). Discussing on the same topic, Gray in *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Gray, 2009) argues that while research projects which fit into the group of the Big Science have a considerable budget for the software and other IT solutions (more than 25%), the ‘small sciences’ have limited capability when it comes to Laboratory Information Management System (LIMS) solutions. Lord in “*From Data Deluge to Data Curation”* (Lord, et al., 2004) describe three main activities in the process of data curation in e-Science. These three main activities are referred as: *data captivation, data storage,* and *data exploration techniques*. This research-work will target mainly the data captivation activities, although the modelling technique influences the data storage layer and is a strong factor in the data exploration techniques. While the management of data in general includes software and hardware components, the hardware aspects are not to be discussed through this research. Digital repositories like EPrints[37], Fedora Commons[38], and DSpace[39] have been extended to accommodate the storage of research data. E-Research infrastructures like Hydra[40], Islandora[41], and eSciDoc are specifically designed for virtual research environments (VRE) and data management. The attention of this work is focused on exploring a novel approach of data modelling and data harvesting from investigation environments in support of the fourth paradigm, data intensive science.

2.4 Linked Data

This section covers basic information on Linked Data (LD) and Semantic Technologies. In Semantic Web terminology, *Linked Data is the term used to describe a method of exposing and connecting data on the Web from different sources* (Webopedia, 2014). The section starts with a discussion and relation of LD with the practice of research data management. Technical details on the underlying technologies of linked data are explained at Section **2.4.2** *Knowledge Representation* and Section **2.4.3**, *Description Logic*. These two sections are important for the understanding of the data modelling methodology followed in the scope of this dissertation and discussed further in Chapter 4.

[37] http://www.eprints.org/uk/ - Digital Repository Software and Services

[38] http://www.fedora-commons.org/ - Fedora Commons is a robust, modular, open source repository system for the management and dissemination of digital content.

[39] http://www.dspace.org/introducing - DSpace is the software used to build open digital repositories

[40] http://projecthydra.org/ - Hydra is an ecosystem of components that lets institutions build and deploy robust and durable digital repositories. The project has originated as University of Alberta.

[41] http://islandora.ca/ - Islandora is an open-source software framework designed to help institutions and organizations and their audiences collaboratively manage, and discover digital assets using a best-practices framework.

The definition of linked data, as the method of exposing and connecting data across different repositories, can be easily related to interoperability, an important feature advocated by researchers and policy makers. Interoperability is defined as *the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units*[42]. With regard to the capability of data communication and data transfer, interoperability is defined by the selected technology used in the data annotation process. This underlying technology should be simple to the users, creators and consumers, and at the same time empower data exchange. There are certain similarities on the requirements of interoperability and the promises of linked data. Both aim to simplify the way we discover, access, integrate and use data.

The linked data vision extends to practices which allow for data publication with intent of improving discovery, findability, classification and integration of these data. As we will see, these features provide numerous benefits in data science activities. The importance of interoperability with regard to research data management and cross discipline collaboration has been stressed by various policy makers and research communities. The report "*Riding the Wave - How Europe can gain from the rising tide of scientific data*" of the High Level Expert Group on Scientific Data, European Commission (High Level Expert Group on Scientific Data, October 2010) presents a vision where the challenges of diverse data formats, people and communities are avoided due to the application of technical, semantic and social features of interoperability. As we can see, there is a tight connection between interoperability and the promises of linked data. This is also confirmed by Heath and Bizer (Heathe, et al., 2011) who argue that linked data evolved to answer questions like:

- *How best to provide access to data so it can be most easily reused?*
- *How to enable the discovery of relevant data within the multitude of available data sets?*
- *How to enable applications to integrate data from large numbers of formerly unknown data sources?*

As we will see, linked data is closely connected with the architecture of the World Wide Web. Similar to the way we connect and consume documents in the web, the linked data recognizes the potential of linking and consuming data in a distributed network of repositories. These repositories are still part of the Web. The differences are related in the *structure* and the *linking capabilities* of the two technologies.

[42] Definition used by ISO/IEC 2382-01, "Information Technology Vocabulary", Fundamental Terms.

Structuring capabilities of HTML and Linked Data

Linked data relies on well-defined data modelling structures. These modelling structures are based on semantic tagging and intend to provide comprehensive descriptive information on the annotated data. In fact, the underlying technology of linked data goes beyond semantic tagging and it is based on practices of artificial intelligence and knowledge representations which we will discuss in the presentation of Semantic Web. The structuring capabilities of technologies of semantic web allow for sophisticated data processing.

The Web on the other side is based on HyperText Markup Language (HTML). HTML's purpose is also the provision of well-defined structures, but these structures are concerned with the representation documents and information created for HTML Clients, such as a Web browser. Therefore, the extent of HTML structuring capabilities is limited to the annotation of *documents* in HTML clients. Metadata such as title, description and keywords in HTML are document oriented. Linked data is concerned with the annotation of *data* and data representations. The structuring capabilities of linked data are related to data elements within an HTML page, or other formats utilized in the Web.

As HTML was natively created to provide interlinking and document annotations, a few attempts have been made to improve HTML structure to inherit some linked data capabilities. The classification of datasets within a HTML document for example, can be tagged with semantic markups. These semantic markups, allow software applications to extract snippets of structured data. With the latest revisions of HTML standards in HTML version 5[43], certain new semantic capabilities are also embedded in the language. These capabilities are still oriented to the annotation of document representation in a client though. Considering the limitation of HTML and representation of datasets in HTML pages, amendments have been made through the introduction of microformats[44]. Microformats are simple markup conventions that enable adding meaningful structure to web content. They allow the publication of structured data within HTML. Microformats provide means to describe types of entities with a limited set of attributes describing these entities. The representation of relationship between entities is also limited. Their usage is aimed to improve data structuring, but their representations has little support for reasoning and classification operations, which as we will see later provide a great value to the semantic web.

As we saw, HTML is a well-structured standard focusing on the structure of Web documents. The Web architecture is very lucrative with regard to the representation of resources, but the HTML focus toward documents limits the representation of resources in

[43] Fifth revision of HTML standards are final and complete since October 2014.

[44] http://microformats.org/

Web Documents. Revisions to the HTML standards have consolidated the focus of HTML to the representation of documents. Data annotation (and data referencing as we will see in the next section) are to be handled through other technologies. These technologies are based on the principles of linked data and improve interoperability. Semantic tagging is the first step toward the semantic annotation and other classification operations in linked data.

Linking Capabilities of HTML and Linked Data

Relationships and connections of documents in HTML are first class citizens, (hence the name *Hypertext* Markup Language). They are based on the HTML *anchor* element *a*. The anchor element can provide a reference to an outgoing link through the *href* attribute. Links in the *href* attribute are interpreted by HTML clients to be followed or traverse directories and retrieve specific documents. This connectivity between documents in HTML has enabled the Web of documents. In a similar fashion, the fundamental idea of linked data is to apply the architecture of the Web to the task of sharing well-structured data in vast repositories. Tim Berners-Lee, acknowledged as the inventor of Web, published a set of fundamentals of linked data. Explaining the linked data principles, he coined four well known rules (Berners-Lee, 2006):

1) *Use URIs as names for things*
2) *Use HTTP URIs so that people can look up those names*
3) *When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)*
4) *Include links to other URIs, so that they can discover more things*

As it can be seen, three of the rules refer to linking capabilities with only one of the principles focused on a common standard. The first principle advocates the use of URIs references to identify any concept. This should not be limited to Web documents (as embedded in the scope by HTML), but also real world objects or abstract concepts (Heathe, et al., 2011). Each of these concepts should be referenced to a URI.

Due to the broad acceptance of the HTTP protocol in the web, the HTTP URI is the recommended referencing mechanism of these concepts. The principle is also known as the *dereferencing* principle. It stresses the importance of providing URIs for each object or concept to be located in the Web.

The third principle of linked data, as advocated by Tim Berners-Lee, is related to a standardization of a common modelling technique. Resource Description Format (RDF), a simple graph based data model is chosen to be the sole model for publishing of structured data on the Web[45]. SPARQL is a query language that is able to retrieve and manipulate data stored in RDF format. It is made a standard by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium, and is recognized as one of the key technologies of the semantic web[46].

The last principle of linked data promotes the use of HTML links to connect not only Web documents, but any type of object or concept. The links in HTML are referenced as document links. Links in linked data are *typed*. Typed links refer to a dependency of the object or concept and their reference. For example, links between an *Investigator* and an *Investigation* can be marked as of the type *runsInvestigation*. These links are recognized as RDF links, to differentiate them from HTML links (Heathe, et al., 2011). Just as HTML links connect any web document in any web page in the WWW, linked data uses RDF links to connect disparate data into a vast fact based repository.

Linked data can play a significant role in interoperability in research data management. The annotation of data with metadata is an important aspect in providing valuable datasets with the potential of exchanging and reusing them. Investigations' metadata can be contextual, describing social and organizational aspects of overall investigations, or technical, describing of the elements, concepts and behaviours recognised during the investigation. With regard to research data management, these *objects*, *concepts* and *behaviours*, together with *contextual information* are part of a universe, which is connected only by pieces of information usually provided in scholarly communication. As a consequence, the results of investigations remain isolated within the group of investigators, within a specific discipline, or within a specific terminology. Linked data presents the paradigm of universe where everything is interconnected. In such data universe, everything has a unique identifier, which is a URI. Metadata dereferencing occurs through objects and concepts that resolve to uniquely identify resources beyond the walls of an institution, or the limitations of a discipline; and whenever different concepts are addressed with different literals or point to different URIs, these concepts can still be aligned with each other. These interconnection capabilities are based on very simple concepts which as we discussed are compatible with the fundaments of Web architecture. A more sophisticated structure of linked data ontologies allows further classification and reasoning operations. These features are discussed in the next section.

[45] RDF is discussed in more depth in Section 3.2 under Data Modelling

[46] "Eleven SPARQL 1.1 Specifications are W3C Recommendations". w3.org. 21/03/2013. Retrieved 12/12/2014

2.4.1 Semantic Web

In *"Semantic Web: A New Form of Web Content that is Meaningful to Computers Will Unleash a Revolution of New Possibilities"* (Berners-Lee, et al., 2001), Berners-Lee et al. present the vision of *Semantic Web*. The article presents the potential of a new extension of the hypertext Web with well-structured data. Websites are to be improved by data services. Smart agents will harvest these data and provide different machine independent operations which will in turn assist the daily life of humans. The Semantic Web's vision is to bring structure to the meaningful content of Web pages, creating an environment where software application can carry out sophisticated tasks[47].

Semantic Web was immediately accepted as the new paradigm of the *new web* as opposed to the classic *syntactic web*[48]. In the syntactic web, the focus was on conveying information to human readers. Minimal descriptive data on the page existed to provide basic information used by search engines to index the document. Interpretation, identification and categorization of web documents are handled by humans (as in internet directories). As the volume of digital data published online increased, attentions rose toward management of valuable resources that needed better organization.

The "*Semantic Web"* article marked a crucial moment in the history of web. Suddenly, the web was no longer considered a thoughtless ocean of informative webpages, but as a big growing repository of valuable data. Nowadays the web is highly service-oriented, fulfilling in part the prediction described in the vision of Semantic Web article. Services such as Intel's Mashery[49] provide a very easy way of providing services by retailers and other sales related companies. API directories such ProgrammableWeb[50] list hundreds of service endpoints which can be used to retrieve and exchange data. Research institutions and companies with modern IT infrastructure are using internal web services to exchange data from different sites and run most of their operations.

The rest of the "Semantic Web" article prediction, the automation of agents who assist human daily operations is still to be exploited. Crucial aspect of this prediction is the

[47] The Semantic Web's vision seemed to complete the missing stage in a word filled with smart robots and artificial intelligence; but the paragraph above is more a technical presentation and based on more concrete technology without the fiction part of the future artificial intelligence envisioned in the Berners-Lee paper.

[48] Syntactic Web is a definition used to describe the current, mostly HTML-based World Wide Web. The term stems from the contrast with syntax, which is the mechanics of a language used to convey information. Semantic Web is related to the term semantics, which is the actual meaning of that information.

[49] http://www.mashery.com/company

[50] http://www.programmableweb.com/

generation of well-defined and annotated data in an interconnected world. This includes data within HTML content in webpages, but also data provided through services or provided by internet repositories. RDF and other standards that are integral part of the Semantic Web paradigm can be used to relate and align data from different repositories to generate knowledge that could be easily extracted and complement with additional information. (Berners-Lee, et al., 2001) predicted the value of well-defined data and their interconnection as opposed to webpage references. Semantic Web related the annotation of data structures found within webpages with the artificial intelligence agents and services that would open a new world in automation of operations.

In order to understand the principle of the Semantic Web it is important to understand the concept of human communication and concepts of knowledge representation. The way we harvest and organize knowledge in information management is still based on abstractions of human operation in accumulating information and organizing knowledge derived by this information. In a conversation among two persons, a combination of ordered symbols and articulations is exchanged. The interpretation of the total messages transferred, is supported by different dynamics such as symbol meaning, prior communications, specific gesticulation or facial expressions. In communication theory the denotation of symbols is referred to as *semantics* and interpretation is known as *pragmatics*. Exchanging data and allowing machines to apply reasoning and interpretation of data is as expected more complex. To enable communication among machines, a syntax that is readable by every machine has to be defined. Information transmitted, should be structured in such a way that it can be used to deduct knowledge. Under these considerations, the metadata modelling should be supported by technology that allows proper communication in an ever increasing digital world. Semantics and interpretation of the underlying concepts represented by metadata should be easily conveyable opening new opportunities for data reuse and interpretation.

2.4.2 Knowledge Representation and Description Logic

Semantic web was envisioned on the premises of artificial intelligence. The underlying technologies of Semantic Web are related to exchange of data and representation of knowledge about specific domains across different services and machines. Discoursing on the underlying technologies of Semantic Web, this section is dedicated to a brief description of Knowledge Representation Systems as a background for the development of Semantic Web Technologies. It is not the focus of this work to dwell into artificial intelligence theories and automata, but rather focus on knowledge representation of metadata in the scope of Semantic Web. Knowledge Representation and Reasoning (KR) are in fact a sub-field of artificial intelligence devoted to the representation of information in a form that computer can utilize it for independent problem solving. Knowledge engineering is the activity of accumulating, maintaining and using information in knowledge base systems in machine readable format.

The product of knowledge engineering can be used in conjunction with automated reasoning tools to produce new knowledge or to prove the consistency of existing knowledge (Farrar, et al., 2010). Knowledge accumulated in knowledge bases can be domain specific. Application of reasoning operations on a knowledge base can derive a set of logic deductions. The sum of the logic deductions derived from application of reasoning operations on a knowledge base, results in a set of statements. These statements form *ontological theory*, or simply an ontology.

The creation of the set of statements that form an ontological theory requires the use of logic as a means of axiomatization. KR as many other disciplines of artificial intelligence, incorporates findings from human psychology about classification and representation of logical facts. Frame language, a technology used for knowledge representation, is based on research which claims that people use stereotypical knowledge to interpret and act in new cognitive situations. Frame languages in KR were used for representing the real world, described as classes, subclasses, slots (data values) with various constraints on possible values. Frames are stored as ontologies of sets and subsets of the frame concepts. They are similar to class hierarchies in object-oriented languages although their fundamental design goals are different. Beside the frame languages, rule-based systems gained traction in KR research. A rule-based system consists of a rule-base (permanent data); an inference engine (process); and a workspace or working memory (temporary data). Knowledge is stored as the total summation of the rules. Rules are of the form "*IF some-condition THEN some-action*". The condition in the rules, test the working memory, e.g. for the presence of certain symbols or patterns of symbols. In many systems, the conditions are expressed logically as conjunctions (occasionally disjunctions) of predicates (Ireson-Paine , 1996). Rules are worthy for representing and utilizing complex logic usually in process-based interactions. They are used classification of concepts by defining systems by a set of logical axioms. It wasn't long before the frame and the rule-based researchers realized that there was synergy between their approaches. Frames are useful for representing the real scenarios with their constrains and possible values. Rules are useful to utilize logic operations. Early attempts to build large ontologies were influenced by a lack of clear definitions. Incorporation of clear semantics and logical formalisms was needed. Languages such as the Ontology Inference Layer (OIL) were based on frame-based systems. Concepts were developed in Description Logic (DL), a family of formal knowledge representation, which is more expressive than propositional logic. OIL was superseded by DAML+OIL, a project of focused on the creation of machine-readable representations for the Web. DAML+OIL concepts are nowadays incorporated in the Web Ontology Language (OWL)[51].

[51] OWL is discussed further in Section 3.2.4.

Members of the OWL family have theoretic formal semantics, and so have strong logical foundations. Description Logics (DLs) are a family of logics that are decidable fragments of first-order logic with attractive and well-understood computational properties. As we will discuss in the later chapters, the use of formalisms based on DL is highly valuable to modelling approaches of research data. The inference indicted by DL influences the search and findability capabilities based on annotated metadata.

2.4.3 Description Logic

In the previous section, we discussed KR systems and the importance of logical axiomatization in providing definitions of the underlying concepts in knowledge-based systems. Axiomatization is defined as *„the process of defining mathematical systems by a set of axioms*" (Dictionary.com, 2015). In the context of this work, axiomatization will refer to the process of defining knowledge-base systems by a set of Description Logic axioms. Axiomatization will be used to generate a formalism on scientific investigation process with focus on the created research data. This section is focused on the presentation of Description Logic and its relation to ontologies. Description Logic is part of a set of formal languages[52] used for knowledge representation. It has emerged from frame-based systems and semantic-networks. As we will see in Chapter 4, Description Logic is the selected language for the formalisms of ontological model development in OWL. In this section will cover basic concepts, syntax and semantics of Description Logic.

As the name narrates, *Description Logic* is a language which makes use of rich semantics to define relation through logical operations. The reasoning capabilities are related to process of deriving the strict logical consequences of assumed premises. The set of reasoning procedures in Description Logic formalism are also influenced by the complexity and decidability. Very expressive Description Logic formalisms are likely to have inference problems of high complexity, or they may even be undecidable. (Baader, et al., 2003) The trade-off between complexity and decidability is an important aspect of application and theoretical research in the field of first order logic (FOL) languages. As opposed to general FOL where logical inferencing is usually undecidable, Description Logic is focused on decidable fragments and nowadays, decidability is conceived as a necessary condition for most formalism in Description Logic (Rudolph, 2011). Expressive Description Logic formalisms may have inference problems of high complexity or might as well be undecidable. Limitation of Description Logic formalisms may be incomplete to represent the important concepts of a domain. As the focus of this work is in modelling a formalism that will be

[52] Formal languages are designed for use in situations in which natural language is unsuitable, as for example in mathematics, logic, or computer programming. The symbols and formulas of such languages stand in precisely specified syntactic and semantic relations to one another. (Dictionary.com, 2015)

applied in real world scenarios, it is important to rely on Description Logic inference which allows a certain level of expressivity and guarantees that any reasoning procedures will always terminate, for positive or for negative answers. Luckily, implementations of Description Logic in ontologies already provide different flavours of Description Logic implementation allowing for applied or theoretical ontological models.

Description Logic represents the knowledge of a specific domain by defining 1) the basic concepts of the domain, this is commonly represented as the *terminology* of a Description Logic and 2) properties of the aforementioned concepts or instances derived by these concepts. These are commonly represented as the attributes of the domain or the description of the domain. The definition of concepts and instances of these concepts allows for classification and structuring. Classification of concepts can be done following sub-concept or super-concept relationships of a terminology. In this case, we create hierarchies of concepts which are useful in inference operations. Classification of instances is related to the concept classification. The set of properties for each instance organizes them in relation to the concept the instance belongs to. In addition, new set of properties for instances may attach new properties to concepts and amass additional facts to the knowledge-base. Description Logic by design has only binary relations and no higher relations (e.g. ternary relations) are allowed. Beside definitions of terminologies and assertions, a Description Logic system offers the ability to process reasoning tasks. Reasoning tasks on the terminology box assess whether a definition is *satisfiable*, or if a definition is more comprehensive (*subsumes*) another. In the same fashion, reasoning tasks in an *ABox* assess whether the set of assertions for an instance is *consistent,* that is the set of properties for an instance qualify the instance to be member of the predefined concept.

The vocabulary of an application domain, or the terminology definition is also referred to as the Terminology Box (*TBox*). The *TBox* consists of axioms about the domain in general in the form of logical sentences. These logical sentences might be referred to as definitions of the terminology. Property definitions of instances of the knowledge-base are part of the Assertion Box (*ABox*). The language for constructing these definitions is a characteristic of each Description Logic system and adheres to specific model-theoretic semantics. Thus, definitions in the *TBox* and *ABox* can be identified with formulae in first-order logic or, in slight extensions of it (Baader, et al., 2003). Relations between instances are defined by binary relations. Sometimes, the *RBox* is considered as another component of the Description Logic system. The *RBox* is considered as a component of the *TBox*. The terms used in scholarly communication differ depending on the discipline of research when addressing the c*oncept*, *individual*, and *roles*. Usually, in the tasks of ontology and knowledge engineering, these terms are used interchangeably with *class*, *instance* and *predicate*.

The simplest definitions in a Description Logic system are based in *atomic concepts* and *atomic roles*. More complex definitions can be built with the assistance logical statements. Basic statements in Description Logic are expressed in the *Attributive Concept Language with Complements* or commonly referred to as *$\mathcal{ACL}$ notation* (Schmidt-Schauß, et al., 1991).

In $\mathcal{ACL}$, the specific class name *T* denotes the concept containing all individuals. Class name ⊥, denotes the empty concept. If *C* and *D* are concepts, then in $\mathcal{ACL}$ we have true the following statements:

- $C \sqcap D$ - the intersection of two concepts is a concept
- $C \sqcup D$ - the union of two concepts is a concept
- $\neg C$ - the complement of a concept is a concept
- ∀R.C - the universal restriction of a concept by a role is a concept
- ∃R.C - the existential restriction of a concept by a role is a concept

In $\mathcal{ACL}$, we can specify subsumptions, e.g. by expressing that every experiment is an investigation via: $Experiment \sqsubseteq Investigation$ (subclass)

Or concept assertion indicating that the individual named *armand* belongs to the set of all *investigators*) such as: *Investigator(armand)*

Other more complex notation can be expressed as well such as: $Instrument \sqcup ComputingDevice \sqsubseteq DigitalInstrument$

As $\mathcal{ACL}$ is a basic notation, additional extensions are made to encompass $\mathcal{ACL}$ and provide more expressive means. In OWL DL, the $\mathcal{SHOIN(D)}$ is used (Horrocks, et al., 2003). As in the case of Description Logic (DL), the capabilities of the notation are included in the name itself. Therefore, $\mathcal{SHOIN(D)}$ stands for:

- $\mathcal{S}$ for the modal logic relation and is an abbreviation for $\mathcal{ACL}$ with transitive roles.
- $\mathcal{H}$ means that role hierarchies are included[53]
- $\mathcal{O}$ means that nominal are included[54]
- $\mathcal{I}$ means that inverse properties are included
- $\mathcal{N}$ means that cardinality restrictions can be used[55]

[53] In RDF this is subproperties - rdfs:subPropertyOf

[54] In OWL this provides support for enumerated classes of object value restrictions - *owl:oneOf, owl:hasValue*

- $(\mathcal{D})$ shows that the use of datatype properties, data values or data types are allowed.

As it was discussed in the description of the terminology boxes, the hierarchical presentation of concepts is an important feature in logical operations. The hierarchy is possible for the terminology concepts and for the roles in a $\mathcal{SHOIN}(\mathcal{D})$ notation. Using role inclusion axioms, $R \sqsubseteq S$ means that R is a sub-role of S. The support for nominals, adds the ability to use data values as second argument of concrete roles. This also allows for cardinality restrictions such as cardinality restrictions:

Eg: Investigation$\sqsubseteq\leq$*hasPrincipalInvestigator* [56]

stating that each investigation has at most one principal which corresponds to *owl:maxCardinality*. The inverse properties, provide some logical assertion by default. In such a case, *hasPrincipalInvestigator* is a role that can relate an *Investigation* with a *Person*. The property *isPrincipalInvestigator*, if declared inverse property of *hasPrincipalInvestigator*, will automatically inherit the restriction to the concepts *Person* and *Investigation.*

While $\mathcal{SHOIN}(\mathcal{D})$ is the foundation notation for Ontology Web Language (OWL), the latest standard OWL 2 relies on another notation termed $\mathcal{SROIQ}$. Decoding the meaning behind $\mathcal{SROIQ}$ would allow us to understand the notation better. In similar to the $\mathcal{SHOIN}(\mathcal{D})$ case, with *S* we denoted $\mathcal{ALC}$ functionalities as the name goes back to the name of a modal logic called *S*. $\mathcal{ALC}$ and *S* could be extended by role hierarchies (obtaining $\mathcal{SH}$ or $\mathcal{ALCH}$) which allowed for simple role inclusions. As it can be seen, $\mathcal{S}$ contains $\mathcal{ALC}$, while $\mathcal{SR}$ subsumes all of $\mathcal{ALC}$, $\mathcal{ALCH}$, $\mathcal{S}$, and $\mathcal{SH}$.

In $\mathcal{SROIQ}$, we are more interested in $\mathcal{SR}$, which represents $\mathcal{ALC}$ extended with all the *RBox* axioms such as reflexivity, antisymmetry (some time referred to as irreflexivity) and role disjointness. Reflexivity is expressed by a role *isIdenticalTo* and in the simplest case, everything is related to itself by this role. Anti-symmetry allows to denote cases where *a* related to *b* via *R* implies that *b* is not related to *a* via *R*. (A classic example of anti-symmetry is the *isParent* role.) Disjointness states that the roles do not share any pair of instances. Example: *hasParent* and *hasChild* are disjoint, while *hasParent* and *hasFather* are not. The *RBox* itself is a set of role characteristics and a role hierarchy.

Table 2: Sequel of logical axioms represented in OWL (Turtle notation) and Description Logic as presented in "Foundations of Description Logics" **(Rudolph, 2011)**

[55] In other terms, the N notation allows counting quantifier, or rules such as: *there exists at least k elements that satisfy property X.* In Owl these are special cases and expressed by *owl:cardinality, owl:maxCardinality*

[56] We will use this notation later on Chapter 4 on providing the definition for some of the entities in the formalized ontology

Axim Type	Turtle Notation	DL paraphrase
Class Equivalence	$[[C]]_C$ owl:equivalentClass $[[D]]_C$.	$C \sqsubseteq D, D \sqsubseteq C$
Class Disjointness	$[[C]]_C$ owl:disjointWith $[[D]]_C$.	$C \sqcap D \sqsubseteq \bot$
Disjoint Classes	[] rdf:type owl:AllDisjointClasses ; owl:members $([[C_1]]_C \ldots [[C_n]]_C)$.	$Ci \sqcap Cj \sqsubseteq \bot$ for all $1 \leq i < j \leq n$
Disjoint Union	$[[C]]_C$ owl:disjointUnionOf $([[C_1]]_C \ldots [[C_n]]_C)$.	$\sqcup_{i<j} C_i \sqsubseteq C$ $C_i \sqcap C_j \sqsubseteq \bot$ for all $1 \leq i < j \leq n$
Property Equivalence	$[[r]]_R$ owl:equivalentProperty $[[s]]_R$.	$r \sqsubseteq s, s \sqsubseteq r$
Disjoint Properties	[] rdf:type owl:AllDisjointProperties ; owl:members $([[r_1]]_R \ldots [[r_n]]_R)$.	$Dis(r_i, r_j))$ for all $1 \leq i < j \leq n$
Inverse Properties	$[[r]]_R$ owl:inverseOf $[[s]]_R$.	$Inv(r) \sqsubseteq s$
Property Domain	$[[r]]_R$ owl:domain $[[C]]_C$.	$\ni r.\mathrm{T} \sqsubseteq C$
Property Range	$[[r]]_R$ owl:range $[[C]]_C$.	$\mathrm{T} \sqsubseteq \forall r.C$
Functional Property	$[[r]]_R$ rdf:type owl:FunctionalProperty .	$\mathrm{T} \sqsubseteq \leq 1r.\mathrm{T}$
Inverse Functional Property	$[[r]]_R$ rdf:type owl:InverseFunctionalProperty .	$\mathrm{T} \sqsubseteq \leq 1Inv(r).\mathrm{T}$
Reflexive Property	$[[r]]_R$ rdf:type owl:ReflexiveProperty .	$\mathrm{T} \sqsubseteq \exists r.\mathrm{Self}$
Irreflexive Property	$[[r]]_R$ rdf:type owl:IrreflexiveProperty .	$\exists r.\mathrm{Self} \sqsubseteq \mathrm{T}$
Symmetric Property	$[[r]]_R$ rdf:type owl:SymmetricProperty .	$Inv(r) \sqsubseteq \mathrm{r}$
Asymmetric Property	$[[r]]_R$ rdf:type owl:AsymmetricProperty .	$Dis(Inv(r), r)$
Transitive Property	$[[r]]_R$ rdf:type owl:TransitiveProperty .	$r\ o\ r \sqsubseteq r$
Different Individuals	[] rdf:type owl:AllDifferent ; owl:members $(\mathrm{a}_1 \ldots \mathrm{a}_n)$.	$a_i \not\approx a_j$ for all $1 \leq i < j \leq n$

Q indicates support for arbitrary qualified number restrictions. It allows for representation of cardinality in the form: ≤nR.C opposed to ≤nR - which is a generalization of cardinality restrictions already present in *SHOIN(D)*. This makes for more concise logical statements, and this is the reason why Description Logic literature usually concerns itself with *SHIQ* rather than *SHIN*.

As mentioned earlier, the presentation of the Description Logic axioms is related to the portability of the formalism to a modelling language. The modelling approach should be used to annotate and allow extraction of information out of research data. In the discussion following in Chapter 3, we will see how the web ontology language (OWL) will be the basis of our formalisation. The foundations of OWL 2 based on the *SROIQ* notation allow the use of Description Logic axioms.

Table 2 presents the portability of a set of axioms to Description Logic and to OWL. (The OWL specification in fact, features some more axiom types than the ones used above but this will be discussed later). Porting a KB expressed in $\mathcal{SROIQ}$ to OWL is done by is done expressing the declarations of the used concept (classes) and role (object property) names and adding a set of definition namespaces, referred to as preamble.

$$[[\mathcal{KB}]] = \text{Preamble} + \text{Dec}(\mathcal{KB}) + \sum_{a€KB}[[\text{a}]] \tag{1}$$

The portability of a set of axioms into an OWL ontology will be discussed further in Chapter 4 where Formula 1 will be used to model an ontological representation in OWL 2. In this section Linked Data was discussed and its benefits to interoperability and research activity. As Linked Data is based on knowledge representation, we discussed briefly Description Logic with focus on $\mathcal{SHOIN(D)}$ and $\mathcal{SROIQ}$. These two notations allow expressivity and guarantee decidability in knowledge base representations. The focus in this section was $\mathcal{SROIQ}$ the underlying notation used for OWL 2 DL, the most recent version of OWL. How axioms expressed in Description Logic can be transferred to OWL and also a formula on the transferability of set of axioms in a knowledge base to an OWL ontology is presented.

Chapter Summary

This chapter's aim is to introduce some fundamental concepts that relate to the contribution of this thesis. The chapter starts with a reference to the field of *Research Data Management* in Section 2.1. This section is focused on the presentation of this practice with focus on digital data. Key concepts such as *digital data*, *provenance* and *contextual information* are discussed and defined in the context of this work. The two following sections propose a discussion on *Scholarly Communication* in Section 2.2 and the newly *Data-Intensive Science* paradigm in Section 2.3. References to these practices are made with the objective of envisioning how the *Research Questions (1.3)* and *Thesis Contributions (1.4)* can influence these two practices respectively. The discussion on these practices continues also in Chapter 5, elaborating the impact and the relation of thesis's contribution to these practices. The chapter's last section is dedicated to presentation of *Linked Data* paradigm and the technology that supports it. The detailed discussion on Linked Data is intentional. This is done to present the vision of the practice. The discussion on the underlying technology stack assists in the quest for a suitable information modelling technology that helps in addressing the thesis research questions.

The selection of a technology for the *Information Modelling* is focus of the discussion of the next chapter.

"Research is formalized curiosity.
It is poking and prying with a purpose."
Zora Neale Hurston

3. Information Modelling and Scientific Investigation

Nowadays scientific instruments and computer simulations are the new standard for scientific investigations in many disciplines. The use of computers in science has changed the way scientists do research, retrieve and disseminate information related to their studies. The motivation of this study is related to information organization and application of new modern practices to improve data annotation and data management activity in computer related research environments. Focal to the research is the development of a feasible data model which allows comprehensive annotation of research data sets. In this chapter we will discuss three main information modelling technologies and select a candidate for further developments.

The chapter is organized in three main sections. Section **3.1** is dedicated to a discussion on a set of requirements that also influence the selection of a specific information modelling technology. The requirements are gathered with respect to the activity of a scientific investigation considering different stages such as modelling research results, live data retrieval with initial (semi-automatic) data annotation, data curation, ingest data management and data interoperability. Section **3.2** is focused on the analysis of three data model technologies that can be used to address the assembled requirements. It provides a discussion on modelling techniques and considerations reflected in selecting a model that can best used to address the aim and research questions of this work. Section **3.3** provides further information on Ontological Data Modelling technology.

3.1 Requirements

The section presents a set of requirements that have been gathered during the last three years based on technical reports, project partners[57], mentors and feedback by colleagues involved in research in disciplines related to information management. These non-functional requirements are to be regarded as criteria that can be used to review the overall functionality of a process, rather than define specific behaviours of an application. The requirements' section serves as a bridge between the research questions and the contributions, which are elaborated in the subsequent chapters.

As stated earlier, the focus of the dissertation is application of new methods with the aim of improving contextual and provenance information conveyed through metadata. In the research questions, the need for good formalization process and a modelling technique is stated. As continuous feedback was gathered to consolidate requirements, it was clear that some of the requests went beyond the scope of formalization in a model. Some of the concerns and requirements of researchers no longer addressed only the acquisition stage in the data life cycle, but they were also tangled with the other stages including versioning, persistent identifiers, authorisation, authentication and data integrity. Traditionally, these processes have been part of repository related operations. Concerns related to these steps have been responsibility of technical staff with regard to policies defined within the preservation plans. With an ever increasing volume of research data, advancements of technology and the ability of digital devices to store data in predefined locations, the data storing process can be altered influencing the data lifecycle flow and repository operations. Reviewing best practices of data management in scholarly communications also indicated different solutions applied to some of these non-functional requirements. In some cases, addressing these requirements was handled in the data modelling process, and sometime they were left to be addressed in later stages such as during data ingest. To clarify on the overall arrangement of data management process, the gathered requirements are described in relation to different layers in a system architecture that is based on a real implementation. The gathered requirements are listed below. For better orientation they are grouped by the *Research Questions* defined in *Section 1.3*.

[57] See Section 4.1 dealing with projects and use cases that influenced this dissertation

Research Question 1: ***How can we model the (finite) environment, entities and relations that are part of an investigation process?***

The question addresses the formalization of a model, which allows the annotation of research data with all relevant entities that influenced the investigation. Luckily these entities are part of a finite (investigation) environment. This environment is composed of technical entities such as devices or conditions, but also social entities such as the investigator, scope of the investigation, project, study, etc. A crucial aspect of the modelling approach should address the relations between all the entities that are part of the environment.

From here we proceed to the first set of requirements.

Requirement 1: *The model shall contain information on entities and relations associated to technical aspects of an investigation.*

A proper description of experiments aiming to facilitate an efficient analysis, annotation and sharing of results is a fundamental portion of scientific activity. Data generated by sensors or other instruments typically lack a lot of important information for the correct understanding and interpretation of experiments. Example: *What instrument was used to capture the data? Was it calibrated? How was it configured? Experiments often require a combination of several instruments (rigs), which may create various artefacts.* These artefacts are related, but these relations remain most often implicit knowledge of the researcher running the investigation. Access to such information is necessary for the reproducibility of investigations and can influence the discovery of datasets across domains.

Requirement 2: *The model shall contain information on entities and relations associated to social aspects of an investigation.*

Another dimension of information that influences the trust in research datasets is related to the social information related to an investigation. Questions such as: *Who conducted the experiment? Who created the data? Which were the initial hypotheses that lead to the generation of these datasets? Who contributed in these investigations? etc.* do not help in the reproducibility of an experiment, but based on the scope and contextual information influence the understanding and the level of trust to the result sets. These data allow for attribution to the researchers involved in the generation of final result-set. Under the social aspects of an investigation additional topics such as: information on the project, institution, study, scholarly communication inspiring and deriving from investigation etc. can be listed.

Requirement 3: *The approach shall assure accessibility features to resources; mentioning access to versioning and persistent identifiers to guarantee longevity access.*

Considering pedigree or lineage (Buneman, et al., 2013) characteristic of provenance, it is to be expected that predecessors of the resource should also be accessible. Features that allow access to prior elements are not new in data management and different approaches already exist for their implementations. Some scholarly communications such as Hartig's *Provenance Information in the Web of Data* (Hartig, 2009), couple the versioning information within the data modelling. In many other scenarios, this challenge is addressed in the repository implementations. The requirement will be discussed with regard to the semantic technology modelling approach charted in this work.

Requirement 4: *The approach shall guarantee that the developed model can be successfully used in an infrastructure*

This requirement is related to the applicability of the model in a practical infrastructure. It is to be shown that the model is not coupled with a specific infrastructure, but can be used with minor amendments across different infrastructures.

Research Question 2: ***How can we use the aforementioned formalization model to simplify the annotation process of research data? Is it possible to automate the process of data annotation and at what extent?***

This question addresses the technological solution that allows for a quality annotation process. The second part of the research question is related to the technology used for the formalization and how to exploit this technology to address some critical questions related to research data dissemination. Some of the requirements under this research question are also related to the Linked Data paradigm.

Requirement 5: *The approach guarantees expressivity and automation*

The process of metadata creation has been traditionally the responsibility of librarians or qualified personnel, who dealt with the final procedures in preserving research data. As we move forward toward preserving massive amounts of research data, several management problems arise. Annotating of these specific data sets is different from annotating documents. The vocabularies used, the models, the components to be described are usually discipline oriented. Considering the volume of the research data generated in scientific investigations, it is impossible to rely on the traditional forms of data annotation. Solutions that the annotation process need to be provided and such solution should work with minimum interaction from the human factor. Yet researchers should have control on the metadata attached to their research results through data curation tools.

Requirement 6: *The approach shall improve discovery capabilities*

Aiming for improved provenance and contextual information means we should also improve the discovery and findability of information. The envisioned approach will provide plentiful information on the technical aspects of investigations. This information should be used to improve findability of the respective datasets. The technology and modelling used should allow for discovery features.

Requirement 7: *The approach will support the concept of interoperability*

The solution should facilitate the process of correlating data across different data repositories, across different disciplines. Interoperability is regarded as very important aspect in research activities nowadays (See (High Level Expert Group on Scientific Data, October 2010)). Solutions that enable such interconnection among disciplines are advantageous to researchers and their importance is continuously stressed by policy makers. It is clear that the separation of discipline related repositories, isolates access to valuable information. Similar restrictions create a problem in finding valuable information. Considering the requirement to improve discovery, interoperability is also a highly important aspect to be discussed in the scope of this work.

Requirement 8: *The approach shall allow for metric information on published data*

Scholarly communication has relied on three main filters to measure the outreach. These three filters are the *peer-review, h-index* and *journals' average citations per article*. Different authors have pointed out that these three filters are one way or another aging (Priem, et al., 2010). Scholars and researchers are heading toward altmetrics and new methods to measure scientific outreach. Altmetrics or the new alternative metrics are based on fetching information on citations in alternative communication channels. Many people relate the new metrics with the increasing significance of online social networks. This claim might be accurate considering the increasing wave of research publications placed in blogging, non-scientific media and other collaboration repositories. Metrics are also related to the sharing of "raw science" like datasets, code, and experimental designs. A solution that support metrics of reach where the citeable unit is a data-set or a group of result sets rather than an article is to be discussed as well.

Requirement 9: *The approach shall guarantee access control support*

Access restrictions and access control is considered a necessity for the publication of research datasets in widely accessible repositories. Despite the movement of open access, there is always a need for a solution which should guarantee protection of the resources from the public access at least for a specific time-frame. The model which will be developed should facilitate or allow integration of an access control mechanisms. A clear solution of addressing this requirement will be documented.

Requirement 10: *The approach shall provide information on licencing*

Licencing information is a very important feature when dealing with data management. The model should offer support to attach clear information about the licencing rights with each instance of information accessed from the service layer of a data repository. The rights to use and re-use the accessible datasets, information on copyright and credits can be provided in different forms at different levels of data aggregations. The model should support access to this information on each data aggregation instance.

As it can be noticed by reviewing the requirements, the quality of the created research data correlates with the complete fulfilment of the requirements. It has to be clarified that the focus of this work is not to provide a full set of requirements that improve quality in research data management. The focus is the improvement of the provenance information, discovery and of interoperability capabilities. The challenge is addressed by using an ontological formalization. The developed method needs *to be generic enough* to allow integration and use of the research data across different application solutions. Although the requirements above are all valid requirements for a good data management, they can be addressed in different components of an application solution. In Chapter **6** I discuss the fulfilment of these requirements.

A set of requirements which are related to a specific modelling formalization are discussed further in Section in **3.2.6**. A full list of requirements can be found in Appendix **B**.

3.2 Information Modelling

Information models have revolutionized data processing and influenced the progress of information technology in last decades. By encapsulating an abstraction layer and hiding technical details, models provide a translation of simplified scenarios from real world problems. Earlier data models have been used since the late 60s to exchange knowledge between humans and computer systems. This section is focused on modelling technique evaluation and an analysis that supports the choice for an ontological modelling technique. The ontological model, as we will see, is flexible with regard to conceptual modelling. It also supports the required data operations that are integral to the research data management process presented in the requirements section.

In the scope of this work three modelling techniques are considered. The three candidate models are selected after a prior analysis. They are considered with regard to their practical use in industry and their potential use in scientific data management as advocated by information technology professionals.

3.2.1 Primer to Modelling

The act of reasoning is typically focused in a specific problem or state of being. The typical analysis process of reasoning deals with the classification of characteristics of a subject and limitation of those characteristics that may influence the considered problem. This simplification approach is a reflexive operation in human reasoning. The same approach has been ported through modelling in different practices. Modelling is perceived as a presentation of the main characteristics that influence and constitute the necessary information-containers to provide valuable information on a problem. The practice tends to be as simple and natural as the reasoning response is to humans. In its simplest form, the practice of modelling is the *simplification of the subject* (Bézivin, et al., 2001). Although the model may be defined as a simplification process, we should bear in mind that the model still needs to present all the attributes of the subject/s that affect the problem and live in the current context. Beside the attributes, the behaviour of the components expressed in a model is also important. While modelling has been understood for a long period as modelling of structures and variables, current analysis focuses on modelling of behaviour operations as well (Chen, et al., 1999). In the context of this work, the following definition will be used to describe a model: A *model is the simplification of a set of subjects within a given context. It describes with accuracy all the properties and behaviour operations that influence the subject's nature*.

Models created with the intention of disseminating knowledge between humans are also referred to as "conceptual models" (Mylopoulos, 1992). Deriving from the term "concept": *an idea of something which is formed by mentally combining all characteristics of a subject*[58]; the conceptual modelling shows how generalization of a problem can be represented. It suppresses non-critical details in order to focus on the main subjects of the problem. A *conceptual model* is essentially a limited representation of a model with humans as the main audience. This same understanding of a conceptual model is also captured by

[58] *Concept*: Dictionary.com Unabridged. Random House, Inc. http://dictionary.reference.com/browse/concept (accessed: July 07, 2014).

Mylopoulos' description where he defines conceptual model as *"[…] the activity of formally describing some aspects of the physical and social world [around us] for purpose of understanding and communication"* (Mylopoulos, 1992). Besides the *communication* of information *among humans*, modelling has inspired a wave of technological implementations. These models are different from conceptual models. They have roots in mathematics and have been crucial to progress in information technology (Mylopoulos, 1992). As it can easily notice, models can be classified based on the targeting audience and the systems of application. In this study, the focus is on information models that find use in information technology.

The quality of these models is subject of evaluation and can be measured with different metrics. Valuation metrics are part of different frameworks that deal with the assessment of the quality of the modelling techniques per sè (such as Lindland, Sindre, and Sølvberg (LSS) (Lindland, et al., 1994), Wand and Weber based on Bunge's ontology (BWW) (Wand, et al., 1990) or the more recent Conceptual Modeling Quality Framework (CMQF) (Nelson, et al., 2012)). The modelling techniques described in the next section are assessed with consideration of our requirements.

From a large variety of modelling techniques, this analysis is looking for the advantages of a specific modelling technique in support of the needs of our data model. I approach the evaluation for the most suitable model by examining the specific layers of modelling representations. These three layers are the *conceptual layer, the logical layer* and *the physical layer*.

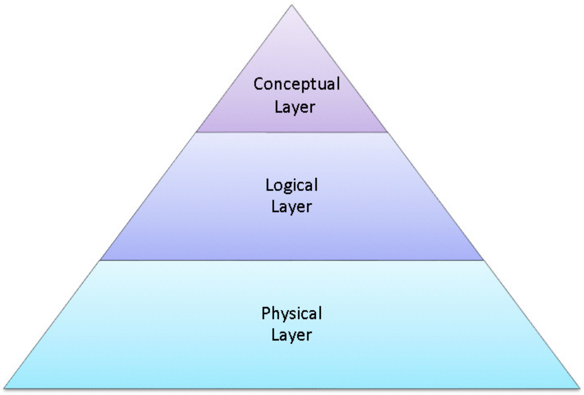


Figure 3: Data modelling layers

In analogy with the conceptual model, a *conceptual layer* is the broad representation of the model. It reflects knowledge on a system, but does not necessarily present the implementation of the information system. The conceptual layer provides an overview of the system being abstracted. This representation is deprived of properties and behaviours. A *conceptual layer* may have some basic identifying concepts or candidate keys. It does not explicitly provide a complete scheme of attributes. Attributes and their set types are logical choices made from a deeper context. This leads us to the second layer as presented in Figure 3, the *logical layer*.

The *logical layer* provides semantic and rational information about the resources in a model. This layer includes information about properties and constrains of properties up to the classification of different entities. The logical layer defines the nature of information that will be used by the physical layer. The *physical layer*, referred as well as the physical data model, specifies implementation details and specific configuration choices for the storage implementation.

If we relate these layers with a practical example, the *conceptual layer* would be similar to a request to have a house build with general information such as the number and functionality of rooms. The *logical layer* would be represented by an architect who defines the exact construct details of the new house and provides other calculations necessary for the engineer. An engineer relates to the *physical layer* that deals with the implementation. Normally, the logical layer defines the complete rationale of a model and the physical model defines the implementations. In specific cases though, the practice has shown that part of these two layers are scattered across the application, business and infrastructure layer in real software solutions. This has been mainly introduced due to limitations in the physical layer, an aspect which we will discuss as an evaluation of the candidate models for our data model design.

3.2.2 Hierarchical Model and XML

Different information modelling practices have emerged to disseminate information across different audiences and disciplines. In an evolving computer world, requirement based on software system representation have influenced these modelling practices as well. The earliest of these techniques were based on hierarchical models that aimed for a simplistic modelling representation. Hierarchical models are materialized in hierarchical databases. They are composed of segments who are arranged in hierarchical order. An example of hierarchical arrangement of data may be seen in XML, which constitutes a repeating history in the developments hierarchical databases.

In a hierarchical model, the topmost segment is known as the root segment. Information relevant to the root segment can be found dispersed among child segments. Figure 4 illustrates the hierarchical model, where the shared hyper-parameter defines a set of properties shared by the third level children. Inference about one group's parameter (level 3) affects inference about another group's parameter. The existence of segments is always restricted by the existence of parents. If we need different derivatives of a segment, which inherit different properties, then we will need a replicated segment. The redundancies of segments lead to inconsistencies in many practical cases.

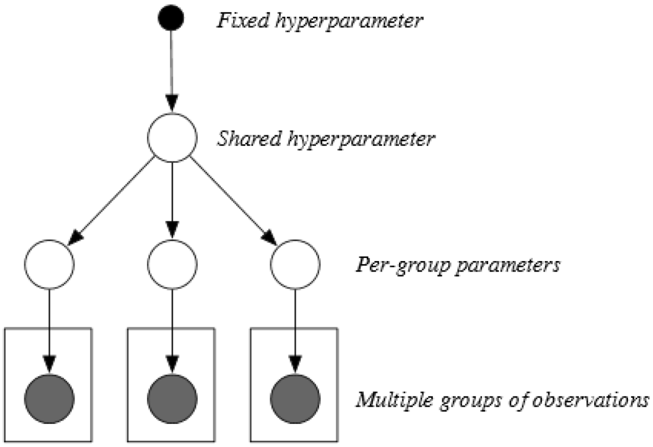


Figure 4: Representation of a hierarchical model (Blei, 2011)

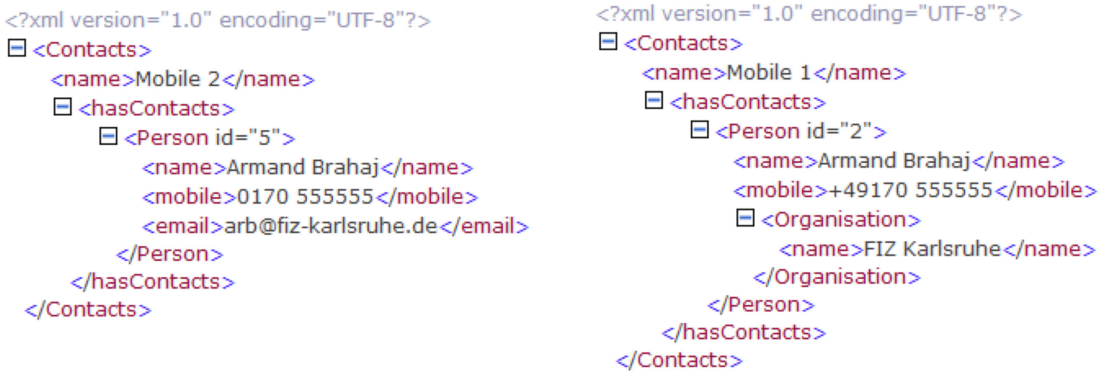
As simple as hierarchical models seem, their implementation in software systems involves operations that are highly influenced by the logical data and physical data independence. Logical data independence refers to the ability to modify the conceptual schema without having alteration in external schemas or application programs. Hierarchical databases such as Information Management System[59] (IMS) supported a level of logical data independence because of the data manipulation language DL/1 (Stonebraker, et al., 2005). Physical data independence refers to the ability to modify the inner schema without having alteration to the conceptual schemas or application programs. Alteration in the internal schema might include new storage devices, use of different data structures, modification of indexes, etc. Hierarchical models faced many challenges in the combination of these two properties and their diffusion in practical implementations.

Early hierarchical models like IMS lost popularity to be revived many years later. Another model, which relates to the hierarchical models, is the Extensible Markup Language (XML). XML is nowadays a very popular language which can be used to organize hierarchical structures (Jörg, 2013). Although categorized under hierarchical models, XML databases adhere also to a category of semi structured data structures (Stonebraker, et al., 2005) due to its newer *schema later* features. Depending on the configurations of the schema, the XML databases can be restrictive or can allow a loose record representation.

[59] IMS is the predecessor of all transaction management and database systems. The system has been maintained and is still a viable product by IBM http://www-01.ibm.com/software/data/ims/

Schema later in contrast to *schema first* suggests that records can be fluid and easy to change. In this interpretation, the schema does not need to be specified in advance. Data instances should be self-describing. This is achieved through tagging of each attribute with metadata that define the meaning of the record (Stonebraker, et al., 2005).

Figure 5 gives the representation of the same "person" entity for a "*contacts*" record. Both XML snippets describe the same entity. Beside the "mobile" field, which has a different representation of the same phone number, the fields are different for each record, although we are using the same concept record type "Person". This representation is possible due to the loose representation allowed by XML with regard to lack of a schema definition. In practical cases, the alignment of the above two records might be a challenge, which can be addressed through semantic mediation.



```
<?xml version="1.0" encoding="UTF-8"?>
⊟ <Contacts>
     <name>Mobile 2</name>
   ⊟ <hasContacts>
      ⊟ <Person id="5">
           <name>Armand Brahaj</name>
           <mobile>0170 555555</mobile>
           <email>arb@fiz-karlsruhe.de</email>
        </Person>
     </hasContacts>
  </Contacts>
```

```
<?xml version="1.0" encoding="UTF-8"?>
⊟ <Contacts>
     <name>Mobile 1</name>
   ⊟ <hasContacts>
      ⊟ <Person id="2">
           <name>Armand Brahaj</name>
           <mobile>+49170 555555</mobile>
         ⊟ <Organisation>
              <name>FIZ Karlsruhe</name>
           </Organisation>
        </Person>
     </hasContacts>
  </Contacts>
```

Figure 5: Record representation in XML Schema later.

The loose features as presented above allow developers to design their own record representation of objects for storing and sharing. Due to this freedom, the XML data model is very useful. XML is applied widely for modelling and exchange of information in many practices in research data management.

In contrast to the loose approach of data modelling in XML, Document Type Definitions (DTDs) and later XML Schema[60] enforced the presentation of records in a well-structured and formatted document. The *schema first* approach deals with validation against a schema of all the records in an XML document. There exist different types of schema performing different types of validations. Structural validation makes certain that XML element and attribute structures meet specific requirements. Data validation looks more closely at the contents of those structures, ensuring that they conform to rules about the nature of information which should be present. Other validation, such as business rules may square

[60] The XML Schema is result of the W3C XML Schema Working Group. More information at the official website http://www.w3.org/XML/Schema

relationship between information and a higher processing level (van der Vlist, 2002). These validations are required to guarantee success of operations in XML databases but also provide check points when models define processes as pipelines of transformation.

Document Type Definition (DTD) defines the document structure with a list of legal elements and attributes. It is limited in definitions and it may be declared inline or as an external reference in an XML file. DTD is based on SGML[61] and the grammar of SGML is not XML itself. XML Schema was introduced as an improvement of the definitions of DTD. It presented a set of powerful features in data modelling and data representation. In contrast to DTD, XML schemas define data-types for elements and attributes; provide support for namespaces which are easily extendable. With regards to communication, XML Schema provides secure data communication. In similar cases the sender could describe the data in a way that receiver will understand, something not possible in case of DTDs.

With the new set of features introduced by XML Schema, the data modelling competences of XML included features which have been flagship of older data modelling techniques. Interlinking and references of records are similar to those in Codasyl[62] data model and Semantic Data Models (SDM)[63]; the set based attribute part of the SDM can also be presented in XML record which can have a set-based attributes; with regard to relational data models XML can have union types, etc.

The simplicity that XML presents in data modelling design and the powerful operations that can be processed through this technology are very tempting. On the other side, as Stonemaker and Hellerstein describe: *"XMLSchema is far and away the most complex data model ever proposed".* The quotation from Stonemaker and Hellerstain continues with: *„It is clearly (that XML) is at the extreme from the relational model on the „Keep it Simple Stupid" scale"* pointing out that XML is also a very complex technology (Stonebraker, et al., 2005). However, XML databases in fact do have a simpler implementation than their predecessors. Serialization of XML can easily be queried with languages such as XPath[64] or XQuery[65].

[61] Standard Generalized Markup Language (SGML), was an enabling technology used in applications such as HTML http://www.w3.org/MarkUp/SGML/ and which was found to have less flexibility in the use with XML.

[62] CODASYL is an acronym for "Conference on Data Systems Languages". This was a consortium formed in 1959 to guide the development of a standard programming language that could be used on many computers. This effort led to the development of COBOL and other standards. With regard to data modelling the Codasyl community was very active and influenced the evolution in the database models through their directed graph data model and their record-at-a-time data manipulation language.

[63] Semantic Data Models will be discussed in the Ontological foundations section in this same chapter.

[64] XPath is a language for addressing parts of an XML document - http://www.w3.org/TR/xpath/

Moreover, the technology has been used successfully in research and industry. XML has been a serious technology candidate for the contribution of this thesis but two more data modelling techniques are considered with regard to the research question and desired outcomes of this work.

3.2.3 Entity-Relationship Model

Entity-relationship (ER) data model also reference as the Entity Relationship modelling (ERM), is a modelling technique that was initially created as an alternative to Codasyl and hierarchical databases in the early 70s (Jörg, 2013). In contrast to the previous models, ERM is based on concepts derived from set and relation theory. Basic to the model are the entities, existing objects or sometimes referred to simply as things, which have attributes to describe their representation and relationships that express interaction between entities. Although initially the ER model was opposed by followers of the Codasyl and hierarchical data models, the technique gained a lot of supporters since it presented the first database model to be described in formal mathematical terms.

In ER, a relation r over a collection of sets (domain values) $D_1, D_2 \ldots D_n$ is a subset of the Cartesian product $D_1 \times D_2 \times \ldots \times D_n$. A relation therefore is a set of *n-tuples* $(d_1, d_2 \ldots d_n)$ where $d_i \in D_i$.

In a concrete example, if we have the following sets:

ResearcherId = {413, 520, 549};
ResearcherName = {Brahaj, Smith, Fisteku};
Project={Nano, Kinetik}

then `r={(413, Brahaj, Nano), (520, Smith, Kinetik), (549, Fisteku, Nano)} is a relation over` *ResearcherId x ResearcherName x Project*.

The relational schema is the relation of the attributes and domain values. Considering a set of attributes $A_1, A_2 \ldots A_n$ associated with domains $D_1, D_2 \ldots D_n$ the relational schema will be the relation R expressed as: $R(A_1{:}\ D_1,\ A_2{:}\ D_2, \ldots, A_n{:}\ D_n)$.

Following our example, the relational schema is a specification of the name and the structure of a relation *Researcher (ResearcherId: Integer, ResearcherName: String, Project: String)*.

[65] XQuery is a flexible language allowing to query a broad spectrum of XML information sources, including both databases and documents. - http://www.w3.org/TR/xquery/

Table 3: Relational Schema and Tuple representation in ER

ResearcherId	ResearcherName	Project	Relational Schema
413	Brahaj	Nano	Tuples
520	Smith	Kinetik	
549	Fisteko	Nano	

Moving from the relational model to a database implementation, a relation instance $r(R)$ of a relation schema can be thought of as a table with *n* columns and a number of rows. Instead of relation-instance we often just say relation. An instance of a database schema thus, is a collection of relations. An element $t \in r\ (R)$ is called a tuple (or row).

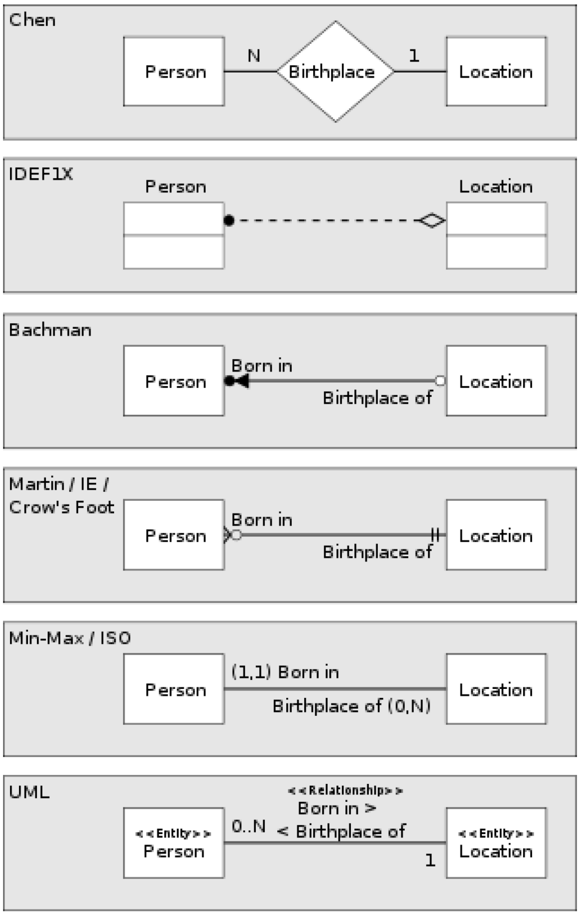


Figure 6: Various representations ER Notations representing the same one to many relationships[66].

One of the criticisms that ERM faced in the early days is the dependency on algebra concepts. To model entities and relation directly in relational schema as above is not an average operation. Diagramming notations were presented to provide a visual guide in representing entities and relations in ER. The most known of these notations are: Chen's

[66] ERD Representation http://commons.wikimedia.org/wiki/File%3AERD_Representation.svg By Ben Thompson (Own work) [Public domain], via Wikimedia Commons

Notation (Chen, 1975), Bachman notation (Bachman, 1969), IDEF1X (Knowledge Based Systems, Inc., 2014), UML (Object Management Group, 2009) etc. An example of the expression of the same entity-relationship in different notations is presented in Figure 6.

The expressivity level of these diagramming notations provided a very accurate overview of entity-relationship. As consequence, a number of approaches emerged to transform a diagram notation, in a relational schema. These approaches make it easy to pass from the basic ER model to a logical model which can be implemented in a relational-database.

Due to this simplicity of transforming a model into a real implementation, the ERM is used extensively in different modelling scenarios, from scholarly teaching to industrial use. On the other side, the relation of the physical data layer and the logical data layer is different from the implementation of hierarchical databases. The logical ER model is developed independently of the physical technology, providing yet another advantage to this model in comparison with the early hierarchical models.

In the previous example, we can see how two entities are linked with each other in the same relational schema. An expected scenario might be the referencing of these entities to other schemas as well. Linking of entities in different relational schemas (tables) is handled through the presence of key (uniquely identifying a tuple) and foreign-keys. In the context of relational databases, a foreign-key is a field in one table that uniquely identifies a tuple in another table. In other words, a foreign key is a column or a combination of columns that is used to establish and enforce a link between two tables of the same database. The database foreign-keys are described because the restriction of linking within the same database structure is a challenge which can especially be noticed in operations which are handled across different systems. These constrains will be discussed later in this chapter in Section 3.2.5 on the assessment of modelling technique.

The set of entity-relationship connection and the relational schema are another example of the “schema-first” restrictions. In this case, everything is presented in a flat table. Relations to other entities (keys) work fine as long as the representation does not change. This has turned to be the weak spot in ER implementations and yet another constrain in data administration. Such a requirement puts the application (RDBMS) before data while an ever-growing avalanche of data, as in our consideration of research focus, requires that data should define the applications.

3.2.4 RDF Model

Resource Description Framework (RDF) (Wood, et al., 2014) was developed with the intention to be a modelling technique for the representation of resources in the World Wide Web. It initially provided a representation solution for the metadata on Web pages like the *title of the page*, *modification date* or the *availability of specific section*s within a website. It did not take long to notice that RDF could also be used to represent any information that can be identified on the Web (Manola, et al., 2004). Due to the impact of World Wide Web and its strong influence on information management, the RDF and the technologies based on it gained a lot of traction in the last decade.

This section is dedicated to Semantic Web or the Web of Data, a paradigm that relies heavily on RDF and advocated by many prominent figures of information technology. The vision of Semantic Web is strongly tangled with the modelling capabilities of RDF.

The RDF model is based on a simple concept: resources on the Web can be described by expressing a set of statements where each statement consists of a *subject*, a *predicate* and an *object*. The subject to be defined in a triple can be identified through a uniform resource identifier (URI) which can be represented as a uniform resource locator (URL). The statement defined by the subject, predicate and an object is called an RDF *triple* and the set of such triples defines an RDF graph.

The RDF represents a *graph* a basic concept studied in graph theory, which is the discipline of mathematical structures used to model pairwise relations between objects. A graph data structure consists of a finite set of ordered pairs, called edges (arcs), which link to certain entities called nodes (or vertices). An *edge (x,y)* points from x to y. The nodes may be part of the graph structure, or may be external entities represented by indices or references. The way nodes are connected to each other through edges defines different types of graphs, which hold different mathematical properties.

A tree as presented in Figure 7, is recognized as an *acyclic simple graph*. An acyclic simple graph is a graph where each vertex has at most one incoming edge. A vertex of degree 1 is called a *leaf*, or *pendant vertex*. An edge adjacent to a leaf is a *leaf edge* or *pendant edge* and so on. As it can be seen, RDF is broader and more inclusive than the hierarchical models expressed as trees and discussed earlier. As we will see later on this chapter, the graph structure and RDF allows more flexibility in modelling than what could have been done with the prior data structures. But first let's have a look at the elements which compose triples.

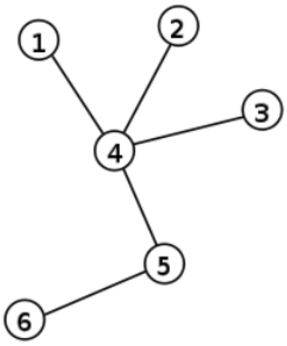


Figure 7: A tree is a simple graph structure recognized as acyclic simple graph. In the figure, a labelled tree with 6 vertices and 5 edges. Nodes 1, 2, 3, and 6 are leaves, while 4 and 5 are internal vertices. If a vertex of the tree is distinguished as root, the tree will be called rooted and will adhere to a hierarchical presentation.

The statement "*Study publication year 2014*", can be regarded as a triple. In this triple, the *Study* is a subject, *publication year* is a predicate and *2014* is the object of the triple. Figure 8 illustrates the graph representation of this triple. It is important to mention that the subject and the predicate, *Study* and *publication year* can be RDF resources themselves. This allows for granular definitions of each resource in a network of graphs. Since each RDF resource has a globally unique identity, the URI for *Study* may be `http://purl.org/net/cosi/#Study` and the URI for a *publication year* might be `http://purl.org/net/cosi/#pubYear` (which might be a reference to `http://purl.org/dc/terms/issued`).

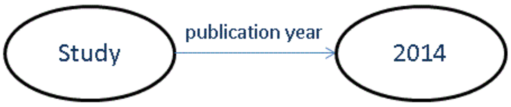


Figure 8: An RDF graph with two nodes (Study and 2014) and vertex connecting them (publication year) forming one triple

As mentioned, the subjects and predicates of a triple are always RDF resources. An object on the other side can be either a resource or a literal value. A relationship is created by having a predicate, which connects two resources. Predicates that connect a resource to a literal value are called *attributes*. In RDF, the relationships to resources are referred to as *object properties* and relations to attributes are referred to as *data properties*.

RDF can be expressed in different ways. Writing RDF graphs can be done through a number of serialization formats. However, different ways of declaring the same graph lead to exactly the same triples, and are thus logically equivalent. The most common form of RDF notation is RDF+XML[67]. This notation inherits has some weaknesses which are influenced by the XML representation. Other well-known formats and language representations are Turtle[68] and N-Triples[69] which are stricter. Below is an example of an RDF graph in Turtle notation.

```
@prefix     :<http://purl.org/net/cosi/> .
:exA        rdf:type    :Experiement .
:exA        :hasSample  "SampleA" .
:hasSample  rdf:type    rdf:Property .
```

As it can be seen from the example above, the RDF is a straightforward representation, focused on the instances and mapping to their types. Since the RDF model is canonical, the RDF data is schema-less. In its base form, RDF has no constraints of range or cardinality. It lacks transitive, inverse and symmetrical properties, which will be described later on. The lack of these properties demonstrates that RDF itself has very limited logical support.

The next layer of RDF model is the Resource Description Framework Schema[70] (RDFS). RDFS extends RDF by defining basic classes that represent the concept of subjects, objects and predicates. These features allow statement definitions about classes of things, and types of relationship.

The modelling capabilities of RDF are enhanced by the schema of RDFS. The schema provides means to represent complex relations such as subclasses, ranges or domains. The relation of resources in these subclasses or domains can be inferred through tools that can perform reasoning tasks based on RDFS. A *reasoner*, or also referred to as *reasoning engine* or *rules engine*, is a software able to infer logical consequences from a set of asserted facts or truisms. These asserted facts are enforced by the presence of the RDF schema. There exist a number of different reasoning engines that can comprehend the RDFS semantics and increase the number of triples based on the relations. For instance, the triples "*exA is an Experiment*" and *"Experiment rdfs:subClassOf Investigation"* imply another triple *"exA is an Investigation"* as in the example below. A group of triples defined by RDF and RDFs create a knowledge graph, which can be queried by SPARQL[71], the query language for RDF.

[67] RDF/XML Syntax Specification - http://www.w3.org/TR/REC-rdf-syntax/

[68] Turtle is a textual syntax for RDF which allows an RDF graph to be completely written in a compact and natural text form. Turtle provides levels of compatibility with the N-Triples (see below) format as well as the triple pattern syntax of the SPARQL. - http://www.w3.org/TR/turtle/

[69] N-Triples are a line-based, plain text format for encoding an RDF graph. - http://www.w3.org/TR/n-triples/

[70] RDF Schema is an extension of the basic RDF vocabulary. - http://www.w3.org/TR/rdf-schema/

[71] SPARQL 1.1 Overview. Query Language for RDF - http://www.w3.org/TR/sparql11-overview/

```
@prefix : <https://www.escidoc.org/ontologies/csmo/> .
:exA    rdf:type          :Experiment .
:Experiement     rdfs:subClassOf    :Investigation .
# After reasoning
: exA    rdf:type    : Investigation .
```

The next layer of expressivity is the Web Ontology Language[72] (OWL). In contrast to RDFS, OWL provides a larger vocabulary of properties and classes to be used. It includes all properties from RDFS such as *rdfs:type, rdfs:domain*, and *rdfs:subPropertyOf.* In addition, OWL adds other classes and properties such as *owl:sameAs* that can be used to align entities among different knowledge-bases. (ex: *ChemiDB:H2O owl:sameAs CityA:Water*)[73]. A list of properties of RDF, RDFS and OWL is presented in Table 4.

Table 4 Property differences between OWL, RDFS and RDF

OWL Properties	RDFS Properties	RDF Properties
owl:allValuesFrom	rdf:type	rdf:type
owl:backwardCompatibleWith	rdf:value	rdf:value
owl:cardinality	rdf:object	rdf:object
owl:complementOf	rdf:predicate	rdf:predicate
owl:differentFrom	rdf:subject	rdf:subject
owl:disjointWith	rdf:first	rdf:first
owl:distinctMembers	rdf:rest	rdf:rest
owl:equivalentClass	rdf:_1, rdf:_2, ...	rdf:_1, rdf:_2, ...
owl:equivalentProperty	rdfs:range	
owl:hasValue	rdfs:domain	
owl:imports	rdfs:subClassOf	
owl:incompatibleWith	rdfs:subPropertyOf	
owl:intersectionOf	rdfs:label	
owl:inverseOf	rdfs:comment	
owl:maxCardinality	rdfs:member	
owl:minCardinality	rdfs:seeAlso	
owl:oneOf	rdfs:isDefinedBy	
owl:onProperty		
owl:priorVersion		
owl:sameAs		
owl:someValuesFrom		

[72] OWL Web Ontology Language Overview - http://www.w3.org/TR/owl2-overview/

[73] Owl:SameAs and the concept aligning capabilities are yet another interesting metric in this study evaluation. It allows interconnection of information among different repositories.

owl:unionOf owl:versionInfo		

OWL adds the semantic richness to formal knowledge representation of RDF, by providing the mean to define the components of triples using proper computable first order Description Logic. While in RDF you cannot check the correctness of statements, OWL provides means to verify their correctness. The logical operations are based on an open world assumption where every fact is considered true unless defined otherwise. Among other things OWL guarantees complex automated reasoning and inference which allow intelligent operations on research data management such as navigation, discovery and data alignment.

There are three different versions of OWL, whose differences lay in the Description Logic constrains used in their implementation. These versions are OWL Lite, OWL DL and OWL Full. Each of these sublanguages is a syntactic extension of its predecessor.

OWL DL is named due to its foundation on Description Logic. It makes use of a reasoner that operates under Description Logic rules and provides all the OWL language constructs that can be used under certain restrictions in OWL DL.

OWL Lite is a simplified version of OWL DL and to be more flexible and have less constrains. The reasoner on OWL Lite supports simpler classifications.

OWL Full has different semantics from OWL Lite and OWL DL. OWL Full allows an ontology to augment the meaning of the pre-defined vocabularies. Unfortunately, OWL Full is considered to be undecidable. Therefore no reasoning software is able to perform complete reasoning for it and its applications are usually academic and out of the scope of this work.

As it can be seen OWL relies heavily on the reasoner and depending on the version, it provides means to express complex constructs. OWL can be expressed in RDF in the same way as RDFS. It can also be queried by SPARQL, although it is more rewarding to query an OWL structure with a Description Logic query. An extension of SPARQL is available for OWL under the query language SPARQL-DL[74].

3.2.4.1 Ontologies in Information Science

As RDF could be used to describe any Web resources it did not take long to notice that these resources could be classified into different categories. As expected the categories

[74] SPARQL-DL API - http://www.derivo.de/en/resources/sparql-dl-api/. SPARQL-DL query engine is settled on top of the OWL API. The library is fully aligned with the OWL 2 standard.

defined elements which shared the same properties. The inverse was also true in many cases. Elements that shared the same properties could be classified into the same category and a parallelism could be drawn between enhanced logical operations and a philosophy branch that studies objects and properties. A correspondence of the classification process was made to the ontology recognized as "*the branch of metaphysics that studies the nature of existence or being as such*"[75].

In information technology, an ontology is *an explicit specification of conceptualizations* (Gruber, 1993) and usually described based on the interpretation of Mario Bunge in his works "*Treatise on basic philosophy: Ontology I: the furniture of the world"* (Bunge, et al., 1977) and *"Treatise on Basic Philosophy (Volume 4): Ontology II, A World of Systems"* (Bunge, et al., 1979). The use of Bunge's interpretation has also been a foundation for the Bunge-Wand-Weber Ontology (Wand, et al., 1990) (Jörg, 2013), which as mentioned earlier is also used as a framework to assess the quality of the modelling techniques.

As Bunge describes, the world is composed of things that have *properties*. These properties can be *mutual* to several things, like the fact that being a child is a mutual property to a father and an offspring, or *intrinsic* such as the name of a person. Substantial properties are referred to as *emergent properties* (like the reasoning is a property of the brain). This leads to the notion of a *composite thing* which owns at least one emergent property. *Attributes* are defined as characteristics of things assigned by people. Attributes might relate to properties, but the difference is that attributes are properties defined by humans. While all attributes are properties, the opposite is not true.

A set of attribute functions defines a *functional schema*. As the functional schema depends on the considered attributes, there might be different functional schemata for a thing, based on the properties considered. A *state* is defined as the set of functional schemata defining a thing at a specific moment. The alternation of the state of things is defined as an *event*. Expressed in a triple, an event is `{InitialState; FinalState; Transformation}` where the transformation will be the mechanism that affects the change. The state and the events are *constrains* which are defined as *laws*, which in turn are also property of things. Another concept in ontology is the *system*, which is composed of interacting things. A system can be viewed both as a thing and as an aggregate of things which possess emergent properties.

[75] Ontology. (n.d.). Dictionary.com Unabridged. Retrieved July, 2014, from Dictionary.com website: http://dictionary.reference.com/browse/ontology.
A history of the term Ontology can also be found at: "Theory and History of Ontology" - http://www.ontology.co/info.htm created April 13, 2000, Retrieved July 2014

One can easily see that the underlying concepts found in Bunge's description of ontology are all consequent concepts with accurate definition. The nature of RDF framework and the Description Logic operations endorsed by RDFS and especially OWL allow the creation of ontologies as computable knowledge in different domains. In more practical terms, ontologies support the sharing and reuse of formally represented knowledge among information systems. The underlying RDF helps to define a common vocabulary in which this shared knowledge is represented. In the context of this work whenever we will refer to an ontology, we will do so by adhering to the following definition: **Ontology** *is an object definition system that relies on logic axioms to relate classes and properties, assure the correctness of underlying statements and support knowledge gathering via inference.*[76]

RDF modelling, reasoning properties provided by Description Logic and classification properties of ontologies back up the vision of the Web of Data or the Semantic Web. Techniques supporting Web of Data are generally referred to as semantic technologies. With regard to knowledge organization and in the context of this study, the use of ontologies provides a lot of benefits in the creation of informative systems, where knowledge is well defined and can be easily related across disciplines and repositories. A comparison of RDF modelling capabilities and semantic technologies versus the other two models, ER and Hierarchical models is given in the next section.

3.2.5 Assessment of Modelling Techniques

This section covers a competitive an analysis of three modelling techniques, Hierarchical Models, Entity Relationship Model and RDF. The analysis will be based on the advantages and limitations of the models in their technical implementation. The requirements expressed in section 3.1 are also considered in the assessment analysis. The assessment of these models is based on their capabilities, advantages and disadvantages with regard to: *Logical and Physical Independence; Data Structure Representation; Linking and Interconnection; Scalability; Practical Use; Intended Use;* and *Support for cross reference checking.*

[76] While information technology is an advancing science and progress, most of the terms used nowadays lack a fixed definition and sometime have different meaning to different groups. The first occurrence of a definition for ontology in information technology is related to a Tom Gruber's appealing to the community of artificial intelligence. Gruber defines ontology as: *"[...]a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set of concept definitions, but more general. And it is a different sense of the word than its use in philosophy*" (Gruber, 1995). The definition of Ontology as well has evolved from the first definition and is still an abstract term to many first time readers.

The valuation is presented in Table 6 with a grading of one to three stars. Due to different underlying concepts in these models, Table 5 presents a glossary of terms used for nodes and arcs across these modelling techniques.

Table 5: Glossary of terms for Nodes and Arcs

Model	**Node**	**Arc**
ER Model	Entity	Relationship
RDF	Node	Property Name
XML	Element	Attribute

Physical layer and dependency to the logical layer

Since analysis is based on implementation of the models, the first metric to be discussed is the physical layer and its dependency from the logical layer. In the case of hierarchical models presented in Section 3.2.2, we see how the physical and logical data independence is difficult to achieve. The problem affected only some of the hierarchical models (such as IMS) and it is addressed by follow-up models and their implementation. As revealed by Stonebraker and Hellerstein (Stonebraker, et al., 2005), ER reflected two lessons learned by the implementation of hierarchical databases.

First lessons dictated that the simpler the data structure, the better are the chances of providing logical data independence. Second, with a high-level language, one can provide a high degree of physical data independence. Consequently, the ER model is based on simple data structures. Access to instances of relations in a relational database is handled by a high level data manipulation language (DML) and the set-at-a-time practice. Set-at-a-time is a processing operation in which commands are executed on specific sets and only after all the set patterns have been detected. Similar attempts have been made to improve the XML access of records such as in (Chen, et al., 2003) where a *NodeSequence* interface is presented with functions that filter, navigate, and transform sequences of nodes simultaneously. In RDF the problem of accessing specific sets has been improved with the SPARQL 1.1 engine. This engine processes data in chunks. The smaller the chunks, closer they are with the set-at-a-time executor (Franz Incorporated, 2014).

With consideration to physical layer and dependency to the logical layer, the ER and RDF Model grade better. They both provide a good independence level without enforcing restrictions on the logical operations or physical implementations. This does not hold true for the hierarchical models like IMS which as we saw earlier couple the two layers introducing implementation constraints.

Data Structure Organization

The models considered in this chapter all provide different organization of the data structure. Hierarchical models rely on tree structures. As mentioned, the tree structures can be very restrictive when used to store complex information. This is mainly related to the challenges of providing sophisticated logical reorganizations of the inner tree structure. An example is the inheritance of properties on the child nodes in hierarchical models. This problem has been addressed through duplication of resources, which inherit different properties. This introduced more complexity in the search ability and the software implementation of the models. The ER model follows a flat structure approach, where information is stored in table representations. ER is supported by a high level data manipulation language as well. Additional findability structures use the power of trees to locate the anticipated information such as in indexes, but these are used only for performance issues and on simplified hierarchical organizations. The ER flat structure is organized after normalization which is the effort to isolate data so that modifications of schemata can be made in separate tables. The data structure choice has highly influenced the performance relational databases. However, the dependency on flat data structures has limited some logical operations which in many cases are left to be addressed in application layers, usually through software implementations outside the database management system (DBMS).

The RDF model in contrast relies in graph structures. Graph structures are inclusive structure for tree data structures and as such benefit from all the advantages of a tree data model. What is more important, the use of the graph model avoids problems mentioned regarding tree structures. Graphs do not necessarily adhere to hierarchical properties and they can link to a different number of nodes.

With regard to XML and internal data structures, another important variance between RDF and XML lies in the representation of information through the internal structure. Considering for a moment the example described previously in Figure 6 in the presentation of XML. One might argue that the information about contacts can be expressed with different XML inner elements structure, each of one representing a different tree. A conceptual model can be expressed in XML in different ways and it is expected to be transformed back to the same conceptual model and contain the same original information. But due to the nature of XML, different tree organizations might be used to express the same information such as in the example below and they might lead to different hierarchical information. This can of course be solved through enforcement of schema but it is not the case of an RDF Triple. Whatever the underlying representation, the RDF statement will always express the same conceptual information.

```
<?xml version="1.0" encoding="UTF-8"?>
<person>
   <email>arb@fiz-karlsruhe.de</email>
   <name>ARB</name>
</person>
```

vs

```
<?xml version="1.0" encoding="UTF-8"?>
<document href="contact">
   <person>ARB</person>
   <email>arb@fiz-karlsruhe.de</email>
</document>
```

As a conclusion, the RDF model and the implementation through graph structures is more advantageous than tree and flat structures. Flat table structures have their advantages when dealing with aspects like performance and scalability, but also limit the logical operations on the internal content. In comparison to tree structures, we saw how graph structures are inclusive structure for trees and benefit from all the advantages of a tree data model. RDF has the upper hand with a very flexible data structure representation. It allows for good content organizations, but also flexible logical operations.

Linking capabilities

The linking capabilities among the considered data model implementations differ significantly as well. Relational databases based on ER provide linking through keys. The use of keys provides a solution to link from tuple to tuple or across tuples in tables. These relations between tuples are though restricted within the scope and within the same database. Linking and aligning two or more records in different relational databases is a challenging process, which in any case cannot be solved by the implementation of the underlying database management system (DBMS). The RDF model on the other side is essentially an opening of the ER model for the Web (Berners-Lee, 1998). Since RDF is composed of simple triples, relationships take an important role in relating these triples. Each relationship is identified by a URI, which in contrast to the ER model, does not have to live within the same object (in ER the identifier is part of the tuple to be related). URI can relate to any other resource in the Web providing a unique feature. XML and XML Schema can also provide relation to resources, but do not have natively the flexibility of URIs in connecting to resources beyond the own document scope.

The linking capabilities of RDF also define some more advantages of this model. The utilization of the ER data models is related to the activity of storing all the information on specific entities within one database-structure adhering to a centralized concept. This constrain is not necessarily part of the RDF data modelling, where one can define an object with some of its properties, and point to another RDF object in another knowledge graph for further properties. This constitutes the decentralized model of the Web of Data where the knowledge on a specific concept can found by merging all the (known) graphs and executing so called federated (SPARQL) queries. This is a great advantage of the Web of Data where information can be obtained from different sources and resulting on a wide knowledge spectrum.

To illustrate the process with a practical example; the *nano* technology is gaining a lot of focus due to the potentials of the usage of *nano* particles. But nano particles may inherit hazardous properties. Different laboratories are examining the potential aspects of these nano particles and different researches find out only a subset of the hazardous properties of nano particles (mainly due to the specific conditions where these particles present the hazardous properties). If one would like to learn of all potential hazardous properties of for example *zinc oxide nanoparticles* (also used daily in sunscreen products), he should query across different repositories (to find out that zinc oxide significantly damages the DNA[77]). Such federated operations are impossible to be handled through the use of ER implementations in DBMS or XML databases across different research repositories. The creation of federated graphs is a distinct attribute of the RDF implementations and an important metrics in this evaluation with regard to research data management.

Scalability

The creation of large knowledge bases by aligning graphs from different repositories has its benefits, which leads us to another evaluation attribute, scalability. While ER implementations in database management systems provide satisfying scalable solutions with regard to the relational operations[78], the same does not hold true for XML and RDF. The scalability of large XML documents decreases due to the nature of the file notation and the redundancy of the contained information. RDF, on the other side, depending on the magnitude of the content, tends to perform better than XML documents due to a simpler internal data notation. Yet the federated queries in RDF are resource expensive. Tests done within the scope of this study show that federated queries do not scale well with regard to the potential they have in relating information from different graph sources.

As it can be seen from the assessment of the three considered models, the RDF data model has some advantages to the other models with consideration of the Logical to *Physical*

[77] Some nanoparticles commonly added to consumer products can significantly damage DNA http://phys.org/news/2014-04-nanoparticles-commonly-added-consumer-products.html

[78] An alternative movement to the implementation of ER systems are the NoSQL data structures which claim to perform better with the price of losing relational features. NoSQL systems are excluded from this evaluation as they provide a minimal logical layer leaving the logical operations to be handled in the application implementations. Such an approach is not desired in our case.

Independence, Data Structure and Linking of resources but has some disadvantages in terms of *Scalability* which might be related to the fact that the implementation of RDF data model is rather young[79].

Table 6: Metrics for assessment of the models techniques

Model Criteria	Hierarchical Models	ER Model	RDF Model
Logical to Physical Independence	Some limitations	Good independence level	Good Independence level
	★	★★★	★★★
Data Structure Organization	Tree hierarchy	Flat tables	Graph
	★	★★	★★★
Linking Capabilities	Limited within the scope of the document	Limited within the scope of a database	No Limitations
	★	★	★★★
Scalability	Limited due to the internal notation and document structure	Scales well (with consideration of the logical operations it supports)	Few limitations in federated search
	★	★★★	★★
Practical use	Widely used in industry and academia, mainly as a serialization format for data exchange	Widely used	Used in industry by the biggest corporations, advocated by information technology researchers
	★★★	★★★	★★★
Intended use	Serialization Format	Object definition in a closed world assumption	Data Model
	★	★★	★★★
Support for cross reference checking	Not part of the native implementation	Not supported in DBMS	First class citizen through federated query
	★	★	★★★

The assessment process is also mapped in Table 6: Metrics for assessment of the models techniques where a 1-3 star evaluation is used with a comment for each modelling

[79] RDF as a data model cannot be considered new anymore. It is a technology which has been advocated for more than 15 years. Although initially embraced by academic communities and research projects, it was not until the last years that the technology finds support by the industry and data management applications. Oracle supported Graph features only with “Oracle Database 10g Release 1 and RDF Semantic Graph” somewhere in 2012. It is for this reason that I refer to the implementation of RDF data model is a rather young.

technique. To review the impact of scalability of a model in practical use, the practical use for each model is also considered. RDF modelling is lately used by the giant of the internet such as Google with its Knowledge Graph[80], Facebook with the Open Graph[81], BBC[82], New York Times[83] and others.

The technology is advocated by information technology prominent figures such as Tim Berners-Lee, the director of the World Wide Web Consortium (W3C). It is focus of intensive research in these years, as it can be witnessed by the increasing number of conferences and journals on Semantic Technologies. The use of relational databases is known to be the core of many information technology organizations in industry and governments. XML, on the other side, has also a widespread use especially due to its use in data exchange communication.

Regarding requirements presented in Section 3.1, I conclude that the use of specific models will influence directly choices under *Requirement 4* through *7*. While *Requirement 4* and *5* are related to a baseline qualification for the selected model, *Requirement 6* and *7* are related to added value of the model. The requirements and their relation to the considered models are discussed briefly below with an emphasis on RDF modelling.

Requirement 4 is related to a successful implementation of the model in an infrastructure. As it was argued, all the three considered models have a certain level of independency between the logical and physical layer. This independency permits the implementation of each model in any infrastructure with minor restrictions. RDF modelling does comply with this requirement.

Requirement 5 is related to the annotation process and the flexibility of the selected model in the process. All the three candidate modelling techniques do comply with this requirement. RDF as well uses the triples, which can be considered as the simplest form of annotating a resource, and as such it conforms perfectly to this requirement.

Requirement 6 is related to the discovery capabilities supported by the selected model. This requirement is at the same time aligned with the subject of the study. An increase in information related to the provenance and contextual information for resources should result

[80] Although Google has always managed to avoid the terms semantic web and RDF, their knowledge graph is based on FreeBase and the full triples can be access from https://developers.google.com/freebase/

[81] Facebook already relies on the power of graphs to relate users and their information within Facebook https://developers.facebook.com/docs/opengraph

[82] BBC Feeds and data provides some information on RDF - http://www.bbc.co.uk/nature/feedsanddata

[83] New York Times has committed to the use of Linked Open Data through their http://data.nytimes.com/ portal

in increasing discovery features. Classification operations of languages based on description logic, such as OWL DL, allow concept alignment across different ontology based knowledge bases. The concept alignment can be used in federated queries to improve discovery capabilities across different repositories.

Requirement 7 is related to the interoperability across repositories of different disciplines. The requirement was influenced and related with the report *"Riding the wave"* (Wood, et al., 2010) that has a strong emphasis on interoperability within disciplines. The vision of the report is also related to the improvement of *"an over archiving multi-disciplinary way of understanding and using data"*. While traditional modelling techniques such as XML or ER have limitation on their native implementation in providing means for the "understanding" of the data, the RDF modelling and the logical operations on top of ontologies are the perfect candidate to fulfil this requisite.

As analysed in this chapter, RDF is the stronger candidate. It provides the same features of the other considered candidates and leverages them further with logical operations, and linking capabilities in the open Web. RDF promises to be the best candidate for the evolving requirements in data management by improving the annotation process, discovery and interoperability among disciplines. It will be the basis of the model designed in the Chapter 4.

3.2.6 Requirements related to the modelling technique

The choice of selecting RDF as the modelling technique, leads to additional requirements. These additional requirements will be used in the assessment process. The requirements are numbered with consecutive numbers to requirements presented in Section 3.1. A full list of all the requirements is documented in Appendix B: Requirements.

Requirement 11: *The formalization shall adhere to linked data (LD) principles for modelling*

The term Linked Data, as presented in Chapter 2, refers to a style of publishing and interlinking structured data on the Web. While formalizing a model, there are a set of principles to be followed which lead to a qualitative process. The linked data community, does not necessary impose publishers to use one schema, but in order to avoid pollution and redundancy it a best practice to maximize reuse of existing vocabularies, URIs, and resources. This requirement should be checked against practices of LD modelling and Green LD (Hoxha, et al., 2011)

Requirement 12: *The formalization shall adhere to Linked Data requirements for publishing*

Tim Berners-Lee has defined a set of rules toward 5 star data publishing practice (Berners-Lee, 2006). Under the star scheme, the quality of publishing data starts by the availability in an open repository (or in the Web). Additional requirements such as machine-readability, non-proprietary format and linked data publishing are the baseline to achieve 5 stars. The approach followed in this works endorses publication of data using semantic technologies and Linked Data.

Although this is an important step, some other requirements should be followed to improve the publishing practice. These requirements include:

- *Completeness of the data*
  partially published data hides valuable information to guarantee trust. It is although acceptable to distinguish in licencing information and provide access control to specific datasets of data
- *Primacy*
  data should be collected at the source, with the highest possible level of granularity. Aggregation and modification can still be used in scholarly articles, but primary data should exist and be accessible

Requirement 13: *Generalizability of the model and portability of the model to other research disciplines*

Generalizability in academic setting refers to the ability for the extension of research findings connected to a specific setting, to other use cases. These requirements are gathered in collaboration with partners and research related to Life Sciences and Natural Sciences. Therefore, the formalization model will address primarily the scientific investigation process in these disciplines. A generalization is possible for other research practices though. The portability of the model to other disciplines such as humanities will be validated.

3.3 Ontology Classification

In this chapter we have discussed the choice of an ontological model for the formalization of scientific investigation. As defined in the presented requirements, the information model that will be developed needs to be generic covering diverse investigation scenarios. It should be comprised of canonical entities and relations. The engineered information model should also allow flexibility in being extended to address more specific needs, depending on the specific research environments. Similar ontologies are referred to as core ontologies. This section is focused on the clarification of disparities on different type of ontologies.

As the use of ontologies has become more and more popular in the last decade, the term is being used in different practical implementations sometime referring to different formalization techniques. The main differences are related to the level of *logical definitions* implied in the ontology, an aspect which will be described in this section. The term *core ontology* has been mentioned in the previous sections but a definition has not been presented yet. This section is dedicated to the presentation of different ontological types used nowadays in practice, and a clarification of the core ontology from a formalization perspective. This clarification will help us understand the different kinds of ontologies avoiding some confusion related to the use of the term in different disciplines in relation to information management activities. This section would also support the choice for developing a core ontology for the formalization of a scientific investigation.

The term ontology has historically inherited some ambiguity in information science and applications in different disciplines relate the term to different knowledge organization schemes. Beside the fact that the ambiguity of the term is linked to the lack of a well-established definition, the practical implementations in different disciplines have also contributed to the confusion around the term. This section describes some different uses of ontology formalizations in different disciplines in attempt to shred light to the definition of ontology. The presentation is important as an intro to the approach chosen for the knowledge modelling technology. It also supports the choice of the selection of OWL-DL and Description Logic used in the ontology development.

Current classifications of ontologies are based on the scope of an ontology and on the level of expressivity of the ontological concepts. Based on the scope of an ontology, there are three main categorizations: *Domain Ontologies, Upper Ontologies* and *Hybrid Ontologies*. Domain ontologies cover concepts that belong to precise realm. They are the most common family of ontologies and sometime rely on *Upper Ontologies* (or Foundation Ontologies) which are more general knowledge base systems providing basic axioms and concepts that are reused by other domain ontologies. Similar Upper Ontologies might be familiar to the reader through DOLCE[84], SUMO[85], WordNet[86] etc. Hybrid ontologies are a mix of the two prior cases and are rarely used.

[84] Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) - http://www.loa.istc.cnr.it/old/DOLCE.html

[85] Suggested Upper Merged Ontology (SUMO) - http://www.adampease.org/OP/

A classification of ontologies can also be related to the level of expressivity and to the breadth of definitions in the internal ontology components such as concepts, properties or relations. Even more important is the level of expressivity related to ontology axioms that are logical sentences used to make explicit assertions about base components. The issue has been part of different studies. Roussey et al. (Roussey, et al., 2011) for example, discussed alterations of ontologies based the language expressivity, on the way the components are defined and on the scope or the domain granularity.

The alternations are a reflective approach depending on specific discipline needs in different practical implementations. With regard to the language expressivity for example, the classification for the concept can be defined through:

- *textual definition* – for example the concept "investigation" is defined by the sentence "part of a scientific procedure in attempt to make a discovery"
- *set of properties* – such as the *purpose*, the *name* of the investigation, etc.
- *logical definition* which might as well be expressed through a Description Logic formula (which we will see in the next section)

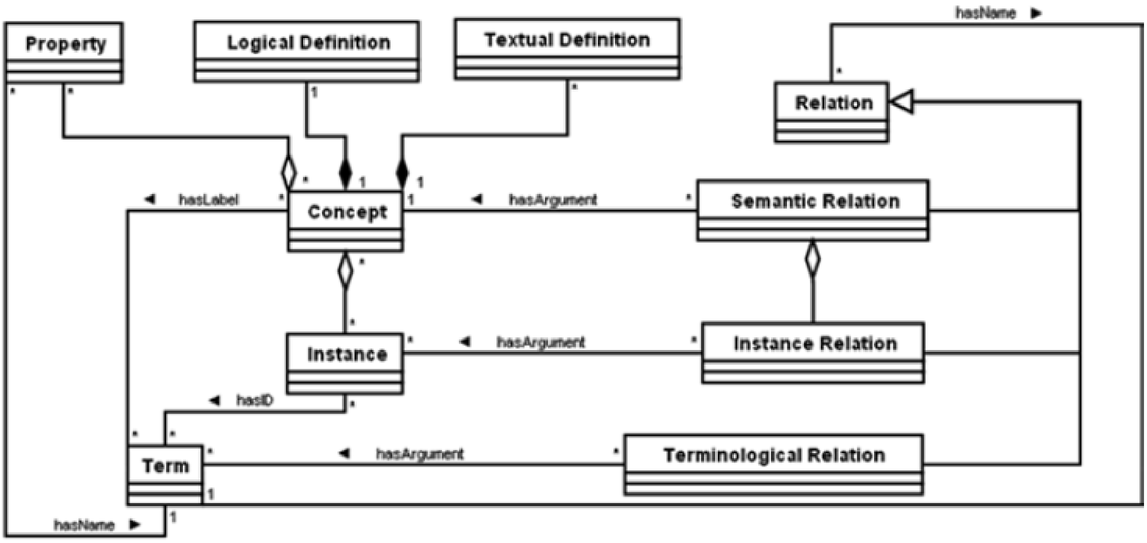


Figure 9: UML Class diagram representing ontology concepts and their relations as presented in **(Roussey, et al., 2011)**

Concepts and objects of an ontology might be described through merely textual definitions, for example in Vocabularies. These concepts can be described by a *set of*

[86] WordNet is a lexical database created initially for the English language. Other ported versions exists for other languages as well - http://wordnet.princeton.edu/

properties as well. In this case, the ontology provides a deeper level of classification based on similar attributes. The classification through sets of *logical definitions* is the strongest and the most correct representation where each concept is seen as the outcome of definitions based on initial axioms.

Figure 9 presents the concepts of an ontology in UML class diagram. As it can be seen, the concept can be represented by one more *properties*, by a *logical* or a *textual definition*. *Semantic Relations* link different concepts influencing the instances of the concept. Relations among instances are also provided through instance relations which might be derived from semantic relations, although not exclusively. An illustration of an instance relation can be the connection of an investigation with at least one site. The fact *an investigation happens in a specific site* is an *Instance Relation*.

Vocabulary relations can also be expressed through a Term which depending on the design of the ontology might be a concept or an instance. The relation among the terms is presented through the Terminological Relation and we will see such relations in the *Appendix C - Vocabulary of Scientific Disciplines*.

While it is advised the use of all three definitions including property, logical and textual definitions most of the ontologies do not necessarily include full formalism of these definitions. Niles and Pease (Niles, et al., 2001) for example mention that due to lack of consensus among computer scientists and philosophers many ontologies are created by librarians and linguists. As they found out, "*those ontologies have typically lacked the formal definitions needed for reasoning and decision making*". The groups mentioned by Niles and Pease include librarians and linguists, each of them interested in *properties* or *textual definitions* of concepts respectively. Related to these two first definitions used in components of an ontology, the first two types of ontologies: *Information Ontologies* and *Terminological Ontologies* can be isolated[87].

Information ontologies are usually used to organize sets of information within the scope of a specific realm through basic property definitions. Similar ontologies have few relations among components and lack semantic and terminological relations (see Figure 9). Such ontologies are easy modifiable and also relaxingly readable by humans. Due to their simplicity they are used not only in describing resources in a realm, but also in other cases such as brainstorming activities and usually represent the first iteration toward more expressive ontologies.

[87] Sometimes, these two ontologies are also grouped under the term „Lightweight Ontologies" (Zaihrayeu, et al., 2007.) (Giunchiglia, et al., 2009) as they have few or limited number of relations.

Terminological ontologies are usually used to describe terms and their relations. These ontologies may have semantic relations and they do have in any case terminological relations (see Figure 9). Their connections can also be defined via relations but they do lack the concept of an instance and instance relations. This is because the expressed entity is itself the instance. The concept can be ambiguous because one concept can be referenced by several terms (in this case *investigation*, *experiment* and *measurements* are synonyms). These ontologies usually are used to present controlled vocabularies, taxonomies and lexical databases.

The classification of interest for us is the *formal ontologies.* These ontologies require a clear definition of components and strict rules which define concepts and relationship. As expected, these ontologies include description of the inner components through *textual*, *property* and *logical definition*. Therefore, the ontology to be built within the scope of this work will adhere to the same guidelines.

Due to their inner elaboration of concepts, the formal ontologies are best suited to be machine readable. They are valuable because in between other benefits they allow reasoning operations. This is usually an advantage of the logical definitions which in ontologies in information systems are usually build with the use of Description Logic. Description Logic (DL) is a family of formal knowledge representation languages where the meaning of the concepts is guaranteed by formal semantics. It is more expressive than propositional logic[88] and has more efficient decision problems than first-order predicate logic. The modelling process covered in the scope of this work is based on formal ontologies, rely on Description Logic and are described in the upcoming section.

[88] Propositional logic, also known as sentential logic and statement logic, is the branch of logic that studies ways of joining and/or modifying entire propositions, statements or sentences to form more complicated propositions, statements or sentences, as well as the logical relationships and properties that are derived from these methods of combining or altering statements. (Internet Encyclopedia of Philosophy, 2014)

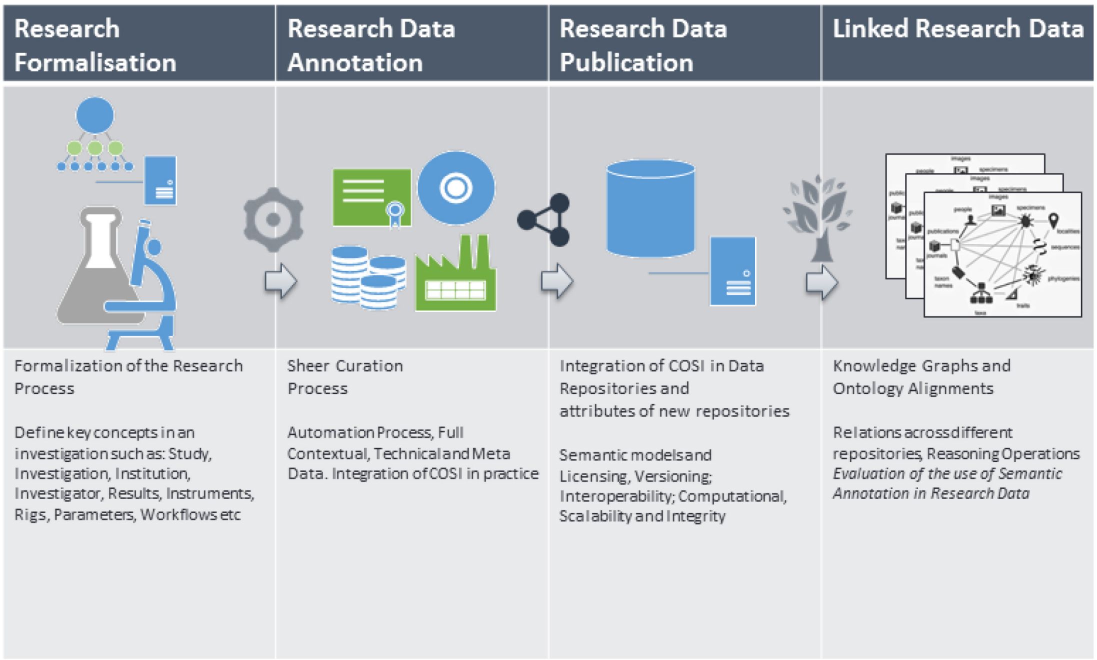
"Science is organized knowledge"
Immanuel Kant (Kant, 1781)

4. Scientific Investigations, Provenance and Contextual Information

The main focus of this chapter is the formalisation of an ontological data model that allows describing the assets employed in the course of a scientific investigation. This ontological data model is formalized based on Description Logic axioms and is expressed in OWL notation. The ontology is named *Core Ontology for Scientific Investigations* (COSI) and its purpose is to allow the disclosure of information covering the entire research process accomplished in the course of a scientific investigation. An ontological model allows flexible logical operations that improve cognitive reasoning, concept alignment, annotations and interoperability of data across repositories and disciplines. These features and the inclusion of all the entities which influence an investigation are the focus of discussion in this chapter. The choice and advantages of an ontological model are discussed earlier in Chapter 3.

As scientific investigations are executed in finite environments, the model will reference and relate all the entities found in similar investigations. The ontology is developed as a core ontology and the intentional use of the term *core* in COSI, relates the focus of the ontology to the set of principal concepts used in scientific investigations. Through a simple meta level of abstraction scientific investigations can be related to a group of basic concepts that form the core of COSI. These concepts can be instantiated by different entity references depending in specific disciplinary scientific investigation. As scientific investigations are executed in finite environments, the model will reference and relate all entities to core concepts. Extending the formalization with discipline specific ontologies is easy and an example is also presented. Documenting all the assets, instruments, parameters, calibrations, provenance and other contextual information related to an investigation can be a resource consuming activity. For this reason, a novel scenario of semi-automated data collecting based on application of COSI is presented. This scenario is focused on the practicability of the model in real life situations.

The applicability of the ontological model in a live virtual research environment and the use of semi-automated curation technique improve the process of storing and publishing investigation datasets. The specifics of the curation technique are addressed in Section 4.3 under Sheer Curation.



Research Formalisation	Research Data Annotation	Research Data Publication	Linked Research Data
Formalization of the Research Process	Sheer Curation Process	Integration of COSI in Data Repositories and attributes of new repositories	Knowledge Graphs and Ontology Alignments
Define key concepts in an investigation such as: Study, Investigation, Institution, Investigator, Results, Instruments, Rigs, Parameters, Workflows etc	Automation Process, Full Contextual, Technical and Meta Data. Integration of COSI in practice	Semantic models and Licensing, Versioning; Interoperability; Computational, Scalability and Integrity	Relations across different repositories, Reasoning Operations *Evaluation of the use of Semantic Annotation in Research Data*

Figure 10: A research workflow followed in this thesis presenting four stages of interest

Section **4.1** is dedicated to the analysis and the methodology that was used in the formalization of COSI. As it also is presented in Figure 10, the second section of the chapter is concerned with the ontological modelling of the key concepts of an investigation. Concrete definition of concepts are discussed in Section **4.2**. This chapter's flow is aligned with the research questions stated in Section **1.3**.

The first part of the Chapter is dedicated to Research Question **1**: *How can we model the (finite) environment, entities and relations which are part of an investigation process?*

In addressing this research question, Section **4.2** is dedicated to the Ontology Formalization. This section is starts by evidencing of the key concepts used in scientific investigation. These entities are used in designing COSI. A specific subsection dedicated to the ontology axiomatization, provides a general overview of logical rule sets of the developed ontology. It provides further clarification on the nature of the ontology model and deals exclusively with the layer of Description Logic embedded in it. The Description Logic operations are beneficial to support scientific interconnectedness and other computational operations whose importance is stressed in Research Question **2**: *How to use the aforementioned formalization model and sheer-curation to simplify the annotation process of result-data? How will this formalization be used to improve the interconnectedness in research activities?*

Section **4.3** Sheer Curation, defines a practical implementation of the developed model in an ever increasing volume of research data. *Sheer curation* is an approach where curation activities are unobtrusively integrated into the normal work flow of a scientific investigation. It improves the rate of data and other digital assets creation. A generic implementation is presented supporting the practice of sheer-curation pointing to the benefit of its use. This section is related to Research Question **3**: *How to automate the publication process of research data in data repositories and still comply with requirements of good scientific publication practices?*

Section **4.4** Data Licencing and Rights Declarations, covers concerns on terms of use for the shared data. These requirements comply to necessities defined under Requirement **10** and **11**. It describes the use of access control mechanism related to a data layer and necessary support from repositories. The topic of information licencing is also considered with regard to information contained in the metadata and also to access control indicators embedded in the model. These two requirements are related to Research Question **1** and **3**.

As already mentioned, the focus of this work is related to scientific investigations covering research disciplines from structural sciences, a subset of Life Sciences and Natural Sciences. Life Sciences, Natural Sciences and other reference to scientific disciplines are used by following the classification of the DFG as published in *DFG Classification of Subject Area, Review Board, Research Area and Scientific Discipline* (DFG, 2008). The model is developed based on analysis and interviews of experts involved in the aforementioned disciplines. The portability of the model to other disciplines such as humanities is also possible. The ability to port COSI to other disciplines is discussed in Section 6.4, *COSI Portability to Other Scientific Disciplines*.

4.1 Methodology followed in the formalization of the ontology

This section is dedicated to the analysis and the methodology that was used in the formalization of COSI. The section starts with a reference to the methodology chosen to model the ontology. In parallel, references are made to reports, technical papers and feedback by domain experts that were used as an analysis for the current version of COSI. The modelling process is presented as iteration through the chosen modelling methodology with references to historical change in the iterations that lead to the actual state of the formalization.

In Chapter 3, discussing on the information modelling, a reference was made to three layers that are to be considered in the modelling process. These layers were the *conceptual layer, logical layer* and *physical layer*. The conceptual layer is accountant for an overview of the model, while the second layer defines more granular properties for the entities that are contained in the information model. The third layer is related to the implementation of the model in a practical information management system. In order to build an information model, an analysis that covers all three layers is needed. This section is focused on the methodology and the workflow used to define the first two data modelling layers that resulted in COSI formalisation. The third data modelling layer, the physical layer is discussed in the upcoming sections from 4.1.2 to 4.1.4.

Although there are different methodologies that can be followed in modelling an ontology (see (Asunción, et al., 2004) (Grüninger, et al., 1995) (Uschold, et al., 1995)), in modelling COSI, a methodology that is similar to a software engineering process was chosen. A typical software engineering workflow contains the following stages *Requirements, Analysis, Design, Implementation* and *Tests*. Deriving from the Unified Software Development Process (UP) (Jacobson, et al., 1999), a widely used standard in software engineering, UPON (De Nicola, et al., 2009) is an ontology building methodology capitalizing on the experience of UP. UPON methodology is based on *Workflows* that contain the aforementioned stages: *Requirements, Analysis, Design, Implementation* and *Tests*.

Requirements and Analysis provide the necessary information required for the conceptual layer. The design stage is related to the logical layer, or a concrete expression of the model in a specific modelling language. A standalone workflow process in UP defines a method for the creation of a specific model version. As new requirements and analysis are communicated or retrieved, changes on the original modelling version are needed. Therefore, iterative versions become a necessity. For this reason, a set of *phases* (see Figure 11) is also defined, with each phase defining a new version of the ontology.

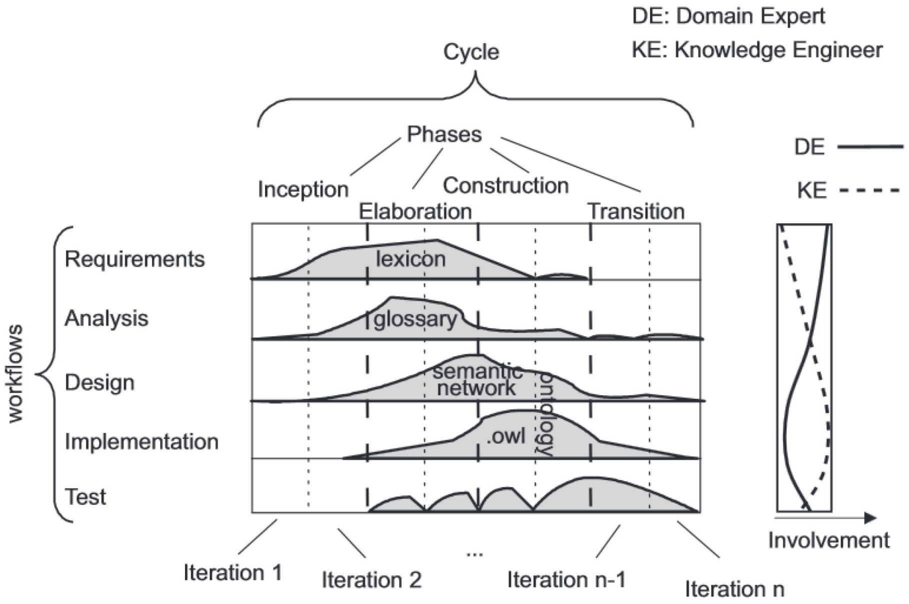


Figure 11 The UPON Framework

In a software engineering development, each phase is seen as a transition through the following steps: *Inception, Elaboration, Construction* and *Transition*. The *n*-set of iterations through phases define a release cycle for the ontology or different versions of it. As one can easily foresee, the early phases of the ontology building are more focused on requirements or analysis tasks. They usually lead to a prototype ontology that is used in the next iteration.

4.1.1 Specification

The *inception* phase, also referred as the *specification* phase, is focused mainly on the process of capturing (new) requirements that will affect the conceptual analysis of the model (see Figure 11 The UPON Framework). The aim of this phase is to allow the modeller have a full insight on the problem being addressed. Capturing and documenting the ontology requirements includes activities such as identification of the *intended usage, use cases and motivation scenarios*. (Decisions on the formality of the ontology are determined during the inception phase. (Uschold, 1996) defines four categories of formality for an ontology: *highly informal, structured informal, semi-formal* and *rigorously formal*. The formality level ranges from expression in natural language, up to classifications in meticulously defined terms. As already discussed in 2.4 Linked Data and 3.2.4 RDF Model, the formality level for the ontology is already defined to rigorously formal, to allow full support for machine processing beside human interactions).

Table 7: Competency Question

Iteration 0 Extracted from *Networking Nanotechnology-Resources for Scientific Education and Research with BW-eLabs* (Jeschke, et al., 2009)	Iteration 1 Based on concepts defined in *Metadatenkonzept für dynamische Daten - BW-eLabs Report* (Razum, et al., 2012) and further elaborated in *Ontological Formalization of Scientific Experiments Based on Core Scientific Metadata Model* (Brahaj, 2012)	Iteration 2 *Core Ontology for Scientific Investigation*
• What is the nature of the experiment being run?	• What is the Topic of the Investigation? • What Programme/study is the investigation related to? • What keywords do the Investigation have? • Who is the Investigator of an Investigation?	•What is the Topic of Investigations? •What is the Programme that the investigation runs? •What is the Hypothesis for the investigation? What are the Hypothesis Statements (acceptance/rejection)? •What is the type of the investigation? •What specific procedure the investigation has? •What institutes are collaborating to the running of the investigation? •What publication motivated the investigation? •What Academic Discipline (Field of Study) does the Investigation have? •What previous investigation preceded the actual investigation?
•What primary data (volatile data) are derived from the current experiments?	•What data are generated from an Investigation?	•What raw-data are generated from an Investigation? •What conclusions did a result have?
•What informal documents and documentations (protocols, notes) are made along the scientific workflow?	•What Publication Derived from the investigation? •What Samples derived from an Investigation?	•What Publication Derived from the investigation? •What Samples derived from an Investigation?
•What is the preservation and access information related to the life cycle of data?		•What versions does a result-set have? •What persistent identifier does a result-set have? •Until when should a resource be conserved? •What copyright information is attached to the data? •What licence is attached to the data?
	•Who has authorisation on an investigation? •Who was authorisation on a dataset?	•Who has authorisation on an investigation? •Who was authorisation on a dataset?

In defining the requirements for an ontology, (Sure, et al., 2002) point out that an ontology should describe the following set of information: *domain of an ontology, goal of the ontology, description of concepts and instances, use cases* and *application support for the ontology*. For the case at hand in modelling an ontology that allows annotation of artefacts and activities in an investigation, the scope of the ontology, potential use cases and usage scenarios were already envisioned in technical reports. The effort for the formalization of an

Ontology on Scientific Investigations started with the project BW-eLabs[89]. BW-eLabs aim was the advancement of heterogeneous experimental resources (remote and virtual) for sustainable coverage. The model at hand is significantly distant successor of the original model introduced with BW-eLabs. The initial conceptual model (not materialized in an ontological model) was focused on the need to retrieve and store primary data of experiments executed in remote instrumentation services. This activity was envisioned in the report *Networking Nanotechnology-Resources for Scientific Education and Research with BW-eLabs* (Jeschke, et al., 2009) a document that defined the desired outcome as anticipated by project stakeholders. It also constitutes the first set of documented requirements used in modelling the ontology related to a scientific investigation.

In the scope of UPON methodology, the requirements phase deal with the collection of the motivation scenarios and semantic needs related to the knowledge that will be modelled. The requirements usually start with the definition of an extent of the modelling system. In the scope of UPON, defining the extent of an ontology consists in identification of sum of the compulsory concepts to be represented. Storyboards, use cases, competency questions (questions at a conceptual level that the ontology needs to answer) and identification of an application lexicon are crucial part of the process. Normally, the modeller has a nebulous idea of the outcome of the connection of each concept in the model. The nebula is cleared as requirements are clarified or new requirements come. New requirements are addressed in iterations. In the case of BW-eLabs, requirements and outlook of the expected model were documented in technical reports such as (Jeschke, et al., 2009) and (Razum, et al., 2012). Based on these reports, a set of competency questions is extracted and presented in the two first columns of

Table 7: **Competency Question**. The last column represents the set of iterative competency questions related to the formalization of the ontology in hand. A resemblance between the competency questions in the last column can easily be drawn to the Requirements already defined in Chapter 3.

4.1.2 Elaboration

The goal of the elaboration phase is to structure the domain knowledge in the form of a conceptual model. The *elaboration* phase is more related to the analysis and definition of proper glossary, entities and attributes that belong to the model. Elaboration as a phase is also important in the consolidation of requirements; therefore, it will affect new versions of the

[89] See as well “Use Cases and Applications” in this same section

ontology as new requirements are presented. The elaboration phase is connected to the analysis step in the workflow. During the *analysis* process, further structuring and refinement of the ontology requirements are delivered. Special focus is dedicated to the consolidation of a *reference lexicon*, the lexicon that will be used to design the model. The reference lexicon is based on the application lexicon (defined in the requirements), and a domain lexicon. The domain lexicon is based on requirements defined by the domain experts and it may be a reference to concepts defined in already existing models, domain standards or foundation ontologies. References to existing models are also evaluated in order to evaluate if alignment and utilization of primitive entities from other ontological models, such as Foundation Ontologies[90] whenever possible.

The identification of the glossary to be used is related to the re-usability and ontology alignment activities as well. Therefore, for the development of the ontology, the following activities are considered in this phase:

- Identification of key concepts describing the domain knowledge
- Efforts to reuse and align the ontology with similar/existing ontologies
- Structure of the key concepts and their relations in a conceptual model

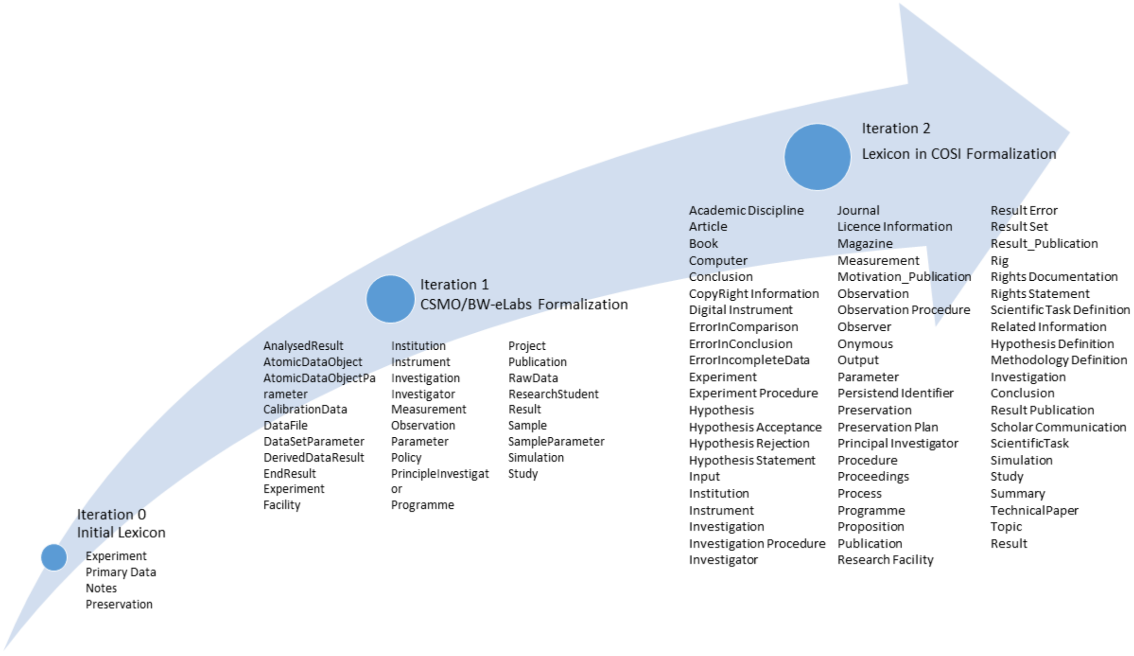


Figure 12: Glossary of terms through different iterations up to the COSI formalization

[90] An analysis on a chosen Foundation Ontology and some other domain specific ontologies used in COSI is presented in Section 4.1.2

Of significant importance on the identification of key concepts describing the domain knowledge was again a technical report by the implementation partners in the aforementioned projects. In *Metadatenkonzept für dynamische Daten - BW-eLabs Report* (Razum, et al., 2012) the key concepts and a foreseen hierarchy of relations were described. The requirements and analysis documented in this report served as background for the modelling effort of a data model, although no ontology model was generated up to this stage (see iteration 0 on Figure 12: **Glossary of terms through different iterations up to the COSI formalization**).

The reference lexicon was validated in iterations with the project partners before iterations releases. The outcome of the analysis process is the definition of a *glossary of terms* to be used. The glossary of terms is the finite sets of terms and their definitions that will be used as entities in the design of the ontology. Figure 12: Glossary of terms through different iterations up to the COSI formalization, shows the change (not necessarily growth) in the terms defined during the analysis phase.

Once the glossary of terms to be used is well defined, review of standards and similar models is made to see if utilization in part or in full of existing ontologies is possible. In the first iterations of BW-eLabs, different existing models have been considered for the task. Some of the considered models were ISO 2146, *Information and documentation - Registry services for libraries and related organizations* (International Organization for Standardization, 2010) and LiLa - *Library of Labs* (Boehringer, et al., 2010). ISO 2146 is an international standard that establishes rules for repositories operating in a network environment. This standard provides information about collections, activities and services needed by libraries and related organizations to manage their collections. Project LiLa on the other side, focuses on a practice for exchange and access to virtual laboratories and remote experiments in laboratories that are remotely controlled through a network connection. Beside these models, similarities to the terms defined in the requirements phase and scope of the model were found in the core Scientific Metadata Model(CSMD) (Matthews, et al., 2010), a model for the representation of scientific study metadata. CSMO/BW-eLabs (see Iteration 1 in Figure 12) made use of a lot of information that was modelled in CSMD. On the other side, COSI formalization, as we will see later on this chapter, has a strong relation to other data models such as PROV-O, for the provenance information, and uses concepts and alignment to primitive entities from other foundation ontologies[91] such as SUMO.

The information gathered through the analysis process results in a reference lexicon or a finite set of terms confirmed by the domain experts and enriched through reference to

[91] An analysis on a chosen Foundation Ontology and some other domain specific ontologies used in COSI is presented in Section 4.1.2

existing domain or foundation ontologies by the knowledge modeller. Table 8, presents an evolution of the ontologies versions and offers some metrics indicating the growth of the glossary, reflected in the total class count.

Table 8: Evolution in metrics of Ontology models that precede COSI

	BW-eLabs Prototype	**CSMO/BW-eLabs**	**COSI**
Class count	38	52	120
Property count	41	57	152
Logical Axioms	197	252	580
Referred ontologies	11	13	20

The *design* step in the UPON workflow completes the information needed in the logical layer. The information that has been gathered in the previous steps is shaped and defined in logical terms. The glossary is composed of entities that in the first cycle iteration may constitute a thesauri ontology, containing a hierarchical organisation of the entities. In consequent iterations, the entities are also extended with object and data properties. The object properties related entities to other entities. Data properties provide reference to predefined variable families. The relation of entities to one-another is an aspect of the *logical layer* in data modelling.

4.1.3 Construction and Transition

The construction phase is related to the implementation, or the encoding of the ontology through a formal ontology language. The c*onstruction phase* is strongly focused on the *design* and *implementation* steps, whereas the last phase, the *transition,* is concerned with the release of ontology versions. Important is also the process of collecting the necessary materials needed for future improvements. This is based on tests and feedback by the domain and community experts. The implementation is also related to ontology editors, and the ontology at hand is encoded partially manually and partially through Protégé (Knublauch, et al., 2004), a well-known tool used in ontology engineering.

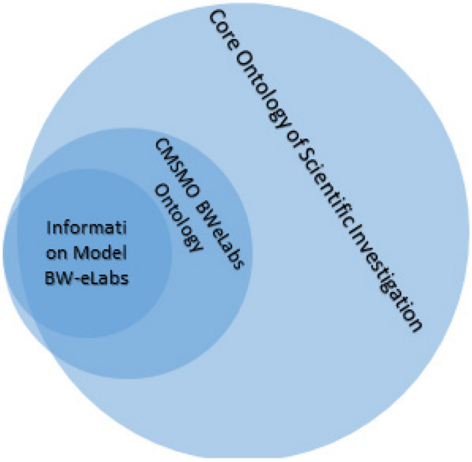

Core Ontology of Scientific Investigation
CMSMO BWeLabs Ontology
Informati on Model BW-eLabs

Figure 13: COSI development as iteration of CSMO/Bw-eLabs

With regard to COSI there are two influencing versions that lead to the current ontology (see Figure 13). A first prototype ontology was released as the "BW-eLabs Ontologie" (Grotendorst, 2011). Although the ontology was result of further requirements and analysis, the ontology per sè did not evolve further, also due to strong ties this ontology had with a repository model, (the first implementation was tightly coupled with the repository application). The ontological model that was used in the BW-eLabs project was the Core of Scientific Modelling Ontology and BW-eLabs Ontology (Brahaj, et al., 2012). CSMO was customized for the BW-eLabs project resulting in CSMO/BW-eLabs, the first ontology iteration in the effort to model an ontology related to capture all information related to a scientific investigation. This ontology was described in the article *Ontological Formalization of Scientific Experiments Based on Core Scientific Metadata Model* (Brahaj, 2012). The publication was important also due to the feedback and acceptance by experts. Beside the requirements defined in the BW-eLabs Project, CSMO relied on Core Scientific Metadata Model (CSMD) (Matthews, et al., 2010). Modelling CSMO/BW-eLabs ontology based on the CSMD model introduced lack of prominent concepts of a scientific investigation and brought the need for another iteration. Most of the primitive classes of CSMO needed improvement to allow for a better interoperability and concept alignment with other existing ontologies as well. Alternation of the primitive classes in CSMO led to new logical definitions of the initial CSMO classes. Iterations and further development of the ontology were implemented in the eSciDoc Generic Browser[92]. Based on the experience of CSMO, partner collaboration and review of state of the art implementations from Laboratory Information Management System (LIMS), improvement over CSMO led to a new incarnation of the model, presenting COSI the ontology at hand.

The construction phase of the ontology relates a lot to the *implementation* and the *tests*. The purpose of the implementation step is the encoding of the ontology in a rigorous, formal language (De Nicola, et al., 2009). This effort is described in more details in Section 4.2.2. The testing is related to an evaluation of the ontology, that is described in Chapter 5 Ontology Evaluation.

4.1.4 Steps involved in the development of COSI

The set of activities, methods and tools involved in the customized methodology are described in the following steps.

Step 1: Analyse the motivation scenario and intended usage

The goal of this step is closely related to the understanding of the requirements of the model. The requirements are related to the information that should be conveyed, as well as to the technology that should best cover the different scenarios. These activities are encapsulated in the analysis of the motivation scenario and intended usage. This analysis is heavily based

[92] See as well the Uses Cases and Applications under the Sheer Curation Section 4.2

on the consultation of technical papers and requirements for applications of the model in a live software.

Step 2: Decide on the formality of the model

The concept of formality is related to degree of formality by which a vocabulary is created and meaning is specified. (Uschold, 1996) identifies four kinds of ontologies stretching from ontologies with no formality requirement at all, to rigorously formal ontologies: The first type of ontologies are those expressed loosely in natural language in the case of *highly informal* ontologies. The second categorization is obtained by adding some degree of structure to the natural language. This results in *structure informal ontologies*. These ontologies are expressed still in a natural language that is well structured resulting in clarity by reducing ambiguity. The third category of formalization is based on an artificial language that is developed with the aim of expressing *semi-formal ontologies*. The highest formalization is used to categorize *rigorously formal ontologies* expressed as meticulously defined terms which formal scemantics, theorems and proofs of such properties as soundness and completeness. In our case, the decision for the formality of the model is based on the necessity of using well structure ontology and is categorized in the last category. Section 2.4 2.4.3 - Description Logic and 3.2.4 - RDF Model provides a rationale for the choice of formality for the model at hand.

Step 3: Gather Competency Questions

Competency questions are questions at a conceptual level that the ontology needs to answer. Gathering competency questions is a process that involves analysis of the requirements and practical usage scenarios. Competency questions express the functional requirements of an application of the model. They help in evaluation whether the ontology fulfils the use cases mentioned in the motivation scenarios.

Table 7: **Competency Question** presents a set of competency questions as they have evolved in different iterations of the modelling effort.

Step 4: Develop Glossary of Terms

The glossary of terms that represents the domain knowledge that needs to be represented in the model to be developed. The motivation scenario and competency questions are crucial to the development of the glossary of terms. The set of terms and their relations are further arranged in a conceptual model. Key entities are further discussed in Section **4.2.2** Entity Definitions.

Step 5: Identify and reuse existing ontologies

Identifying and reusing existing ontologies is not only a quality requirement, it also allows alignment of concepts and greatly improve findability based on classification and inference operations allowed by the underlying description logic. Aim of this step is to identify existing ontologies and consider reusing whenever possible their definitions. Identification of appropriate ontologies is related to finding the most suitable foundation ontology and similar domain or case specific ontologies. These two group of ontologies and concrete reuse of other ontologies is discussed in Section Ontology Preamble 4.2.1.

Step 6: Develop Conceptual Model

The conceptual model is a pre-requisite step in developing an ontology. It is related to the structuring and arrangement of the concepts based on their meanings, classification and relations. A simple structure is a hierarchical organization complemented by additional relations that result in a graph structure, a common structure for ontologies. For the modelling effort at hand, the initial conceptual model documented in the technical report (Razum, et al., 2012) was further elaborated. During this step, the five components classes, relationships, functions, instances and axioms were included in the conceptual model along with their constraints. In different iterations, as the glossary of terms grew bigger, so did the conceptual model. Figure 12: Glossary of terms through different iterations up to the COSI formalization presents the growth of these terms in different iterations of the ontology.

Step 7: Formalization and Encoding

Formalization of an ontology relates to the expression of concepts by use of logical axiomatization. Concepts organized in the conceptual model are described in terms of logical definitions whenever possible. Use of foundation ontologies and other domain ontologies allows "borrowing" definitions from other ontologies. For the ontology at hand, a definition described in description logic language is expressed for at least one of the entities described (see Section 4.2.2.1 Investigation for an example of a definition in description logic). Following the formalization efforts, the next expected step is encoding the ontology in a formal ontology language. The purpose of ontology implementation is to develop an artefact that encodes the domain knowledge in a format that is understandable by different type of users, including humans and machines. The ontology at hand is encoded in OWL2 and expressed as Turtle Notation at Appendix C.

In the next section, the ontology is presented with focus on the main reused definitions, entity definitions and axiomatization.

4.2 Core Ontology for Scientific Investigations (COSI)

In this section, Core Ontology for Scientific Investigations (COSI) is described as a formal ontology recommended to be used as an information model for the preservation, mediation and interchange of research data gathered during the execution of scientific investigations. COSI is designed to allow for the documentation of a wide range of provenance and contextual information about such investigations. The application of COSI enables information exchange and integration between diverse sources of research data. The formalisation of the ontology is based on Description Logic through the OWL2 language. It empowers semantic definitions and reasoning inference that in return allow for interconnection and reuse of data, may they be within an institution's repository or federated repositories on the Internet. The ontology is light and can easily be extended for more specific use cases.

4.2.1 Ontology Preamble

In Chapter 2, a presentation of the main topics that influence this work was offered. Presenting Linked Data and the next Web of Data, a reference was made to *knowledge representation* and *Description Logic* as foundations for the development of ontological representations. In Chapter 3, we discussed the definition of an ontology as an object definition system that provides precise formulations of entities, properties and relations in a specific domain. The formalization of an ontology is established following Formula 1.

$$[[\mathcal{KB}]] = \text{Preamble} + \text{Dec}(\mathcal{KB}) + \sum\nolimits_{a€KB}[[\text{a}]] \qquad (1)$$

The first part of the formula is related to the Preamble. The preamble is a set of namespace definitions, or in other words, references to formalizations that may be used within the ontology. The initial Preamble in COSI is presented in Formula 1.1.

$$\text{Preamble} = \begin{cases} @prefix\ owl\colon\ < http{:}//www.w3.org/2002/07/owl > . \\ @prefix\ rdf\colon\ < http{:}//www.w3.org/1999/02/22-rdf-syntax-ns > . \\ @prefix\ xsd\colon\ < http{:}//www.w3.org/2001/XMLSchema > . \\ @prefix\ rdfs\colon\ < http{:}//www.w3.org/2000/01/rdf-schema > . \end{cases} \qquad (1.1)$$

The preamble in 1.1 contains references to other formalisms such as RDF, XSD etc. Integration of other formalisms in one's own ontology is a recommended practice. Some researchers have also attempted to use the term *Green Linked Data*, see *"Toward Green Linked Data"* (Hoxha, et al., 2011), as a concept of reusing as many formalisations and prior-rulesets instead of redefining over and over new rulesets. Integration of already existing ontologies affects as well the relation (ontology matching) and alignment capabilities of concepts across ontologies. In this section we discuss the ontologies referenced in the preamble of COSI together with a group of *Foundation Ontologies*.

The second component in Formula 1, the declaration of the knowledge base is expressed through the formula:

$$\text{Dec}(\mathcal{KB}) = \sum A \in N_C(KB) + \sum r \in N_R(KB) \qquad (1.2)$$

where A denotes class types. Example: *A rdf:type owl :Class* . Lower *r* stands for object properties. Example *r rdf:type owl:ObjectProperty* .

The last component of the Formula 1 is expressed as a summation $\sum_{a \in KB}[[\text{a}]]$ where *a* denote axioms allowed to be used in the underlying logic of the ontology. Axioms provide explicit logical assertions about classes, individuals and properties. We will discuss the axioms as we cover the definitions of our entities and their relations in the ontology.

The preamble and namespaces of the ontology are essential to any ontology formalism. The preamble defines inclusion and relation to supplementary concepts and ontologies. Normally, a single ontology may contain elements and attributes that are defined and used by different communities. The reusability of ontologies is of course defined by the granularity and abstraction level of the formalisation. The more complex an ontology, the more difficult it is to be reused due to redundancy and possible unneeded complexity. Incentives for the utilization of namespaces are *practicality, modularity and reusability*. The preamble or the definition of namespaces in an ontological model shows that certain elements will be inherited by a specific formalisation, referenced in a specific namespace. Considering that each element in an ontology is reference by an URI, the use of namespaces allows shorter notation of the elements and this is practical when expressing or querying ontologies. As we progress in this chapter, we will see that some of these namespace definitions are always found in the declaration header of an ontology in OWL. They allow reusability of elements and properties from *RDF, RDFS* and *XML*.

The use of elements and properties from other formalisations require that the document constructs and contains universal names whose scope extends beyond the containing document. *A namespace is a collection of names, identified by a URI reference, which is used in documents as element types and attribute names*. The use of the term namespaces in this document is the same to the common definition of *Namespaces in XML 1.0*[93]. Names from namespaces may appear as qualified names that contain a single colon

[93] http://www.w3.org/TR/REC-xml-names/

separating the name into a namespace prefix and a local part. The prefix, which is mapped to a URI reference, selects a namespace for use. The combination of the universally managed URI namespace and the document's own namespace produces unique identifiers. The modularization empowered by namespaces is an important feature in building ontologies.

From an ontology engineering perspective, the construction of an ontology should be based on the combination of self-contained, independent and reusable knowledge components. The linked data community highlights continuously the importance of taxonomies reuse of already existing ontologies. The reuse of other ontologies in order to derive concept and properties taxonomies is also important because it improves concept alignments across different ontologies. For example, the entity *Investigator* can easily be placed hierarchically under a *Person* entity. In order for other systems to derive that the Person entity denotes a real human being, a full set of properties has to be defined. Luckily, other ontologies already have defined this entity. Listing *Investigator* under the *Person* entity of a well-known ontology saves a lot of resources in entity definition. At the same time, linking to terms and roles defined in well-known ontologies will improve logical alignment operations which are highly important in federated queries, or search across different knowledge bases with different ontological description. External ontologies to be reused may be domain specific ontologies or foundation ontologies.

In order to provide a full definition of the namespaces provided in the preamble of the ontology, a discussion on the upper ontologies is presented and referenced as *Foundation Ontologies*. The definition of the knowledge base and all the axioms will continue in Section 4.1.2, *Entity Definitions*.

Foundation Ontologies

Foundation ontology, also known as a *top-level ontology* or *upper ontology*, are a special ontology focusing on universal description of concepts. These ontologies are usually high level and domain independent. Foundation ontologies provide a set of base definitions for top level concepts and properties. Normally, domain specific ontologies extend and derive concepts and properties from foundation ontologies. In the definition of COSI, a foundation ontology is also used to provide some basic entity definitions in the formalisation. Foundation ontologies aim to define broad concepts that are meta, generic, abstract and sometimes, philosophical. Important characteristic of an upper ontology is the focus on a very broad semantic interoperability between a large number of ontologies which are accessible "*below*" this upper ontology.

Table 9: Preliminary evaluation criteria for the considered upper ontologies

Criteria\Ontologies	**BFO**	**DOLCE**	**SUMO**	**OpenCYC**
Licensing	BSD New	No Licence / No Restrictions	Free/GNU	Apache License
Structure	KIF Ported to OWL	KIF Ported to OWL	KIF Ported to OWL	CYC Ported to OWL
Maturity	Used in many practical implementation	Mainly used in Academia	Used in many practical implementation	Used in many practical implementation

Although there are several ontologies recognized as foundation ontologies, in this section four of them are considered based on their *practical use*, *domain scope* and *references in scholarly communication*. To come up with a corpus of foundation ontologies for evaluation, a total of ten foundation ontologies based on Google Scholar results[94] were initially selected for evaluation. Out of these initial ten foundation ontologies, four ontologies were filtered based on h-index ranking and the pre-evaluation criteria presented in Table 9. The analysis below presents the rationale in selecting a foundation ontology for COSI.

The evaluated upper ontologies are *Basic Formal Ontology* (BFO) (BFO - Basic Formal Ontology, 2015), *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE) (Masolo, et al., 2003), *Suggested Upper Merged Ontology* (SUMO) (Pease, 2002) and *OpenCYC* (Cycorp, 2015). In the preliminary evaluation, I considered the following metrics: *Licencing*, *Structure* and *Maturity* or *record of use for the ontology in industry and academia*. The *licencing* metric relates to flexibility of creating an ontology with no complications on reuse in scholar research or commercial products as a derivation of a foundation ontology. The *structure* metric is used to filter out foundation ontologies modelled in a logic language not compatible with the selected technology (OWL and RDF Serialization). And finally, the *maturity* is yet another index to filter the most prominent foundation ontologies based on their usage in practical implementations. The maturity was evaluated based on academic and practical usage of the ontologies.

Of these four-leading upper-level ontologies, BFO (Smith, et al., 2010) is claimed to be closely tailored to the needs of scientist users (Smith, et al., 2010). For this reason, BFO is the first candidate to be evaluated as an upper ontology for the formalisation of COSI.

BFO was developed as a small representational artefact with the mission of providing an upper ontology. The focus of the ontology, as its authors claim is *to support the integration*

[94] Google Scholar is a service by Google in support of scholarly communications - http://scholar.google.com/

of multiple heterogeneous ontologies developed for purposes of scientific research (Smith, et al., 2010). In fact, the focus is so specific for BFO, that it seems it has application only in development of ontologies in the biomedical domain. BFO is well-known and used extensively as a foundation ontology for *The Open Biological and Biomedical Ontologies*, also known as the OBO Foundry[95].

Entities formalized in BFO are split into *continuant* and *occurrent* (see Figure 14). Even though many concepts can be defined from these two entity types, (such as *Investigation* deriving from *Process*, an *occurrent entity*), the knowledge-base concepts contained in BFO are very limited. In order to provide the plethora of concepts or contextual information needed to envisage a scientific research activity, a lot of derivative definitions are needed to be created. Relying on prior definition of entities is the first reason why an upper ontology is considered. Therefore, BFO was found unfeasible for the task of supporting the provenance and the contextual information for an investigation activity in the course of this work.

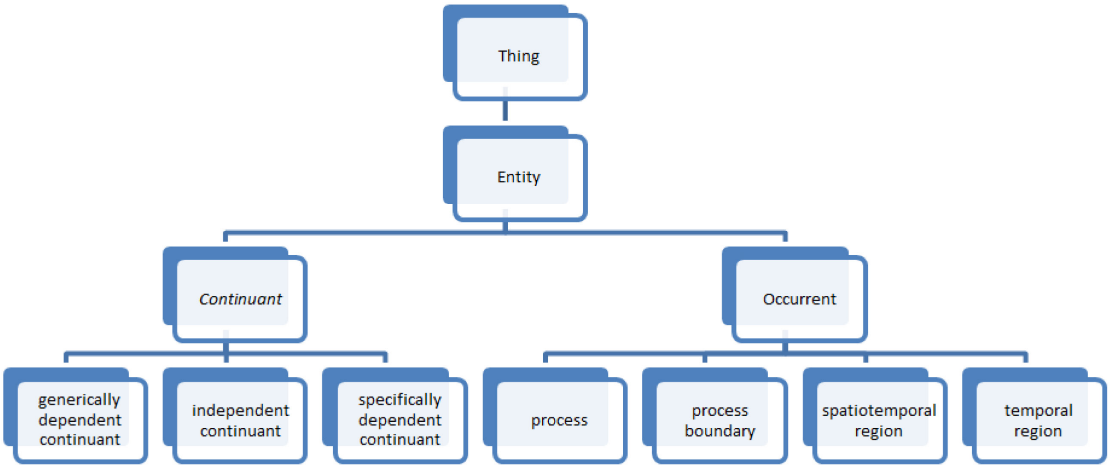


Figure 14: Top Level categories in BFO

As claimed by the authors, BFO is "*focused on the task of providing a genuine upper ontology to be used in support of domain ontologies developed for scientific research*" (Basic Formal Ontology (BFO), 2015). In fact, BFO's value is to be exploited as a supporting ontology for investigation parameters and results, especially (and in practice, exclusively) to the use in the biomedicine domain. BFO is improbable to be used as an upper ontology in a broader concept of scientific research, due to its very high level knowledge representation entities. For this reason, it is ruled out as a potential foundation candidate for the object of this study.

[95] The Open Biological and Biomedical Ontologies - http://www.obofoundry.org/

BFO grew out of two other well-known foundation ontologies, DOLCE (Gangemi, et al., 2002) and SUMO (Niles, et al., 2003). In similar fashion to BFO, DOLCE also has a very abstract and limited set of concept definitions. From a simple observation, DOLCE is also very similar to BFO, containing the basic entities, but in a different hierarchical organization. For example, the *continuant* and *occurrent* entities found in BFO are analogous to *endurant* and *perdurant* entities in DOLCE. These two entities are placed under *spatio-temporal-particular* entity together with *quality* (See Figure 15). Beside the *spatio-temporal-particular,* DOLCE has also included an *Abstract* entity, mainly due to the fact that its aim is to capture the ontological categories underlying natural language and human common sense. This was not the case with BFO, which targeted entities that are scientifically well defined. Abstract concepts seem to have been omitted from BFO as they are non-material entities or metaphysical concepts.

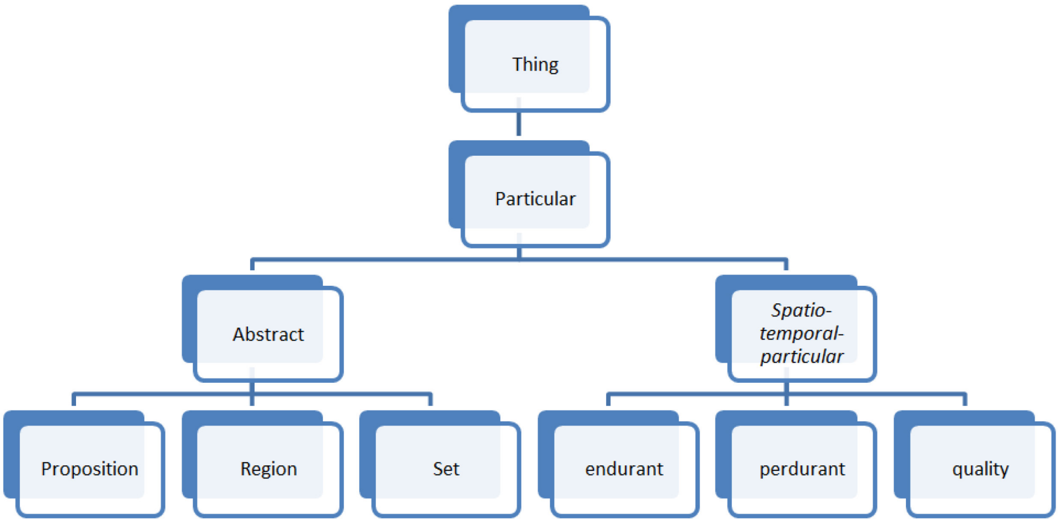


Figure 15: A subset of top level categories in DOLCE as derived by OWL representation

Although DOLCE has a focus on what it calls “linguistic and cognitive engineering” (Gangemi, et al., 2002), in practical terms, this upper ontology inherits the same abstraction characteristics as BFO. With this consideration, DOLCE does not provide significant advantages over BFO in providing support for the development of reference ontologies used to define a scientific investigation.

SUMO and its domain ontologies (Niles, et al., 2001) form one of the largest formal public ontology in existence today. SUMO was developed to facilitate data interoperability, information search and retrieval, automated inference, and natural language processing. In order to make use of the necessary philosophical and linguistic concepts, SUMO makes use of the knowledge base contained in WordNet (WordNet, 2010), a lexical database of English language. WordNet includes nouns, verbs, adjectives and adverbs from the English language.

(WordNet dataset for other languages exist as well.) These linguistic elements are grouped into sets of cognitive synonyms, each expressing a distinct concept.

Harvesting the English dictionary concepts from WordNet has propelled SUMO to be one of the largest ontological knowledge bases at the moment. The abundance of concepts has influenced the use of SUMO. Beside WordNet, SUMO is extended with many domain ontologies from Communications, Countries and Regions, Distributed Computing, Economy, Finance, Engineering Components, Geography, Government etc. It is implemented in the first-order logic language SUO-KIF, but a SUMO OWL version exists as well. An advantage of SUMO, with regard to the evaluation of this work, is the focus of the ontology on "pure" representation and reasoning capabilities. From the practicability perspective, many applications, including academic, commercial or governmental make use and refer to SUMO.

The last upper ontology considered is OpenCyc, the free version of Cyc[96]. OpenCyc is developed by Cycorp, a commercial entity. It is a formalised representation of facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. The name "cyc" derives from "encyclopaedia" and the aim of the project is to provide a formalized knowledge-base of the world. The original knowledge base is proprietary (Cyc). OpenCyc is a reduced version of the Cyc knowledge base. In a similar way to the SUMO set of domain specific ontologies, the Cyc knowledge base is divided into thousands of "microtheories" focused on particular domains of knowledge. The commercial Cyc and the Cyc's micro theories contain more than 300,000 concepts and about 3,000,000 assertions. It is claimed that Cyc contains a knowledge based equal to a 900 person-years of effort (Laningham, 2008). Instead, OpenCyc comes with a reduced number of concepts. In the last release, OpenCyc had about 26,000 concepts. The representation language for Cyc is the CycL formal language[97], which is a formal language whose syntax derives from first-order predicate calculus and the programming language of Lisp[98]. The latest release of Cyc allows exporting concepts and assertions from CycL to OWL through an Ontology Exporter. The translation of OpenCyc to OWL ontologies is found to be incomplete though, leading to limited reasoning capabilities.

From the four foundation ontologies evaluated, BFO and DOLCE ontologies are ruled out due to their specific scope of application. SUMO and OpenCyc appear to be better candidates, especially considering the rich set of concepts they contain. While SUMO is

[96] Cycorp Official Website- http://www.cyc.com/

[97] CycL Syntax - http://www.cyc.com/cyc/cycl/syntax

[98] Lisp is a programming language created as a practical mathematical notation for computer programs. It is one of the oldest computer languages (with Fortran ranked as the oldest). Due to the strong mathematical notation relation, Lisp has been revived and used over and over and is still an influencing software developing language.

absolutely free and widely embraced in academia, OpenCyc seems a luring ontology derived by a commercial product. Although it contains a large number of concepts, the OpenCyc is not considered in the development of the ontology for scientific investigation. With the additional concerns over the limitations of OpenCyc representations in OWL, SUMO is selected as a foundation ontology for the implementation of COSI.

Beside the use of a foundation ontology, a number of domain related ontologies provide also a valuable base of concepts. In some cases, these definitions are more accurate with consideration of technical aspects, especially if the ontologies have in common a specific discipline or application industry. As this work is highly related to the provenance of metadata, the definitions of the PROV Data Model (PROV-DM) (Moreau, et al., 2013), a W3C recommendation, is found helpful in defining a set of generalized concepts related to provenance. In more details, the PROV Ontology (PROV-O) (Lebo, et al., 2013), the ontological derivative of PROV-DM is used to borrow some provenance related concepts in COSI.

Relations to digital objects, files, folders, and other references to data objects are inherited from the NEPOMUK Information Element Ontology (Mylka, et al., 2007). Project NEPOMUK is an attempt to provide unified vocabulary for describing native resources available on the desktop-machines. The project has a number of ontological definitions which are directly or indirectly in the scope of COSI. Annotations of the concepts, classes and relations of the entity are expressed by using some classes of the famous Dublin Core Metadata Initiative (The Dublin Core Metadata Initiative (DCMI) , 2012). The Ordered List Ontology Specification (Abdallah, et al., 2010) is used to provide basic concepts and properties for describing ordered lists in a semantic graph.

Beside the basic definitions of *owl, rdf, xsd, rdfs*, the preamble is completed with references to the foundation ontology and the other domain specific ontologies. With regard to the first part of the formula, and the evidence of the necessary ontologies used in COSI, the preamble of the scientific ontology makes use of the following namespaces:

$$\text{Preamble} = \left\{\begin{array}{l}
@prefix : < purl.org/net/cosi/> . \\
@prefix\ owl: < http://www.w3.org/2002/07/owl > . \\
@prefix\ rdf: < http://www.w3.org/1999/02/22 - rdf - syntax - ns > . \\
@prefix\ xsd: < http://www.w3.org/2001/XMLSchema > . \\
@prefix\ rdfs: < http://www.w3.org/2000/01/rdf - schema > . \\
@prefix\ sumo: < http://www.ontologyportal.org/SUMO.owl > . \\
@prefix\ cosi: < https://purl.org/net/cosi/> . \\
@prefix\ dc: < http://purl.org/dc/elements/1.1/> . \\
@prefix\ nao: < http://www.semanticdesktop.org/ontologies/nao > . \\
@prefix\ nfo: < http://www.semanticdesktop.org/ontologies/nfo > . \\
@prefix\ nie: < http://www.semanticdesktop.org/ontologies/nie > . \\
@prefix\ nrl: < http://www.semanticdesktop.org/ontologies/nrl > . \\
@prefix\ olo: < http://purl.org/ontology/olo/core > . \\
@prefix\ prov: < http://www.w3.org/ns/prov - o - 20130430 > . \\
@prefix\ terms: < http://purl.org/dc/terms/> . \\
@base\ < http://purl.org/net/cosi/> .
\end{array}\right.$$

The namespace of the ontology to be defined, is denoted trough the prefix *cosi (*and blank (see first line, indicating a local scope of the definitions). The namespace for our ontology is defined as *http://purl.org/net/cosi#.* The same URL is used as a technical documentation of the ontology. Beside the base namespace, the preamble contains a set of other prefixes which include the necessary modelling formalisms in defining an OWL ontology. As we have seen in Chapter 3, the OWL definitions share concepts with *rdf* - *http://www.w3.org/1999/02/22-rdf-syntax-ns#* and *rdfs* - *http://www.w3.org/2000/01/rdf-schema#.* The *xsd* namespace is included with consideration of the serialization of the ontology in a RDF/XML representation. The preamble is described in Turtle notation.

4.2.2 Entity Definitions

In this section, the fundamental concepts of the ontology are introduced with informal explanations and illustrative examples. As discussed in *Introduction*, the set of fundamental concepts included in the ontology is collected with attention to *provenance* and valuable *contextual information* required to better understand a scientific investigation and the results of a scientific investigation. Each of the information provided in this section, is important with respect to information complementary to results of scientific investigations. As described in the previous chapter, the modelling technique chosen to be used in the course of this work is based on OWL representation. Entities and the relations that define this model are evidenced through a definition and expressed in Turtle notation.

4.2.2.1 Investigation

Investigations are crucial to the contextual information accompanying a final research dataset. The investigation entity serves as a hub and relates a research dataset to technical and social contextual information. In the scope of the modelling activity, the investigation is to be considered a broad concept that can be represented by more concrete entities such as

Experiment, Measurement, Observation and *Simulation*. These sub classes of investigation are discussed as we progress in this section. An investigation is a process of examination, aiming to discover facts or proofs in support of a hypothesis. Hence, for the scope of this work *Investigation* is defined as:

> **Definition 2:** ***Investigation*** *is an examination process in support of a hypothesis. It is part of a systematic study and adheres to a scientific procedure.*

An investigation has a title, a description and is identifiable. It has at least a principal investigator role and an undefined number of investigators. In Definition 2, a set of key concepts are underlined to indicate that they are still to be defined in the scope of this work. These definitions are provided as we progress in this section. The definitions are mapped to concepts defined in the foundation ontologies or are defined in the scope of COSI. The definition of each concept is denoted by the prefix, a colon and class name. In the case of the investigation defined in Definition 2, since the concept is defined in COSI, it will be represented as *cosi:Investigation*.

In Definition 2, the investigation is defined and related to other concepts. The set of relations to other classes is also presented in a tabular representation as in Table 6 for *cosi:Investigation*. The presentation includes the set of properties that relate the concept with other classes. In OWL we distinguish between three main property types. *Object properties* that link classes together; *data properties* that link a class to a literal, for example an integer or a string and *annotation properties* provide a set of metadata for the OWL concepts.

Table 10: Description of Investigation class in COSI

Class	cosi:Investigation
IRI	http://purl.org/net/cosi#Investigation
Equivalent to	sumo:Investigating
Definition	Investigation is an examination process in support of a hypothesis. It is part of a systematic study and adheres to a scientific procedure. An investigation has a title, a description and is identifiable. It has at least a principal investigator role and an undefined number of investigators.

Has Superclass	cosi:Process;
	cosi:hasHypothesis exactly 1 cosi:Hypothesis;
	cosi:hasProcedure some cosi:Procedure;
	cosi:hasInvestigator some Investigator;
	cosi:hasInvestigator min 1 cosi:PrincipalInvestigator;

Subclass by	cosi:Activity; cosi:Onymous

inheritance	
Described by inherited properties	cosi:belongsToStudy; terms:title, terms:description; terms:identifier (Onymous) cosi:hasActivity (Process) prov:endedAtTime, prov:generated, prov:invalidated, prov:qualifiedAssociation, prov:qualifiedCommunication, prov:qualifiedEnd, prov:qualifiedStart, prov:qualifiedUsage, prov:startedAtTimedp, prov:used, prov:wasAssociatedWith, prov:wasEndedBy, prov:wasInformedBy, prov:wasStartedByop (Activity)

As we can see from Table 10, the *Investigation* class is linked to other classes via its own properties, and at the same time it inherits properties from parent classes. In the super classes of *Investigation* we can see some declaration provided in the form of a triple. The declarations such as `cosi:hasInvestigator some cosi:Investigator` is actually an *anonymous classes*. Anonymous classes are used to declare and instantiate a class at the same time. They are like regular classes except that they do not have a name. Anonymous classes are usually considered when they are used once and not repeated. In the previous example, for any member to be classified as a *cosi:Investigation*, the anonymous class definition should be satisfied. In the case of the predicate *some*, the class enforces an existential restriction[99]. An existential restriction needs to be satisfied for each individual of a class. The anonymous class `cosi:hasInvestigator some cosi:Investigator` indicates that at least some investigator need to be present for an investigation to happen. Existential restrictions describe classes of individuals that participate in at least one relationship along a specified property. Existential restrictions may be denoted by the existential quantifier ∋. They are also known as *someValuesFrom* in OWL terminology.

The definition of investigation, and the definition of the anonymous class `cosi:hasInvestigator some cosi:Investigator`, can be descriped in the following notation in Description Logic:

`cosi:Investigation ⊑∋ cosi:hasInvestigator.cosi:Investigaton`

The predicates "*exactly 1*" and "*min 1*" in anonymous declarations `cosi:hasHypothesis exactly 1 cosi:Hypothesis;` and `cosi:hasInvestigator min 1 cosi:PrincipalInvestigator` are respectively *exact cardinality* and *minimum cardinality* restriction[100]. Cardinality restrictions specify the number of relationships that an individual

[99] Existential Quantifiers, together with the Universal Quantifiers are part of Quantifier Restrictions.

[100] Exact cardinality and minimum cardinality are part of the Cardinality Restrictions

must participate in for a given property. In this case, the declaration `cosi:hasHypothesis exactly 1 rdf:Literal` denotes the fact that each *Investigation* needs to have exactly one hypothesis to exist. These two definitions can be represented by the following notations respectively:

$$\sqcap \geq_1 \texttt{cosi:hasInvestigator.cosi:PrincipalInvestigator}$$

$$\sqcap =_1 \texttt{cosi:hasHypothesis.rdfs:Literal}$$

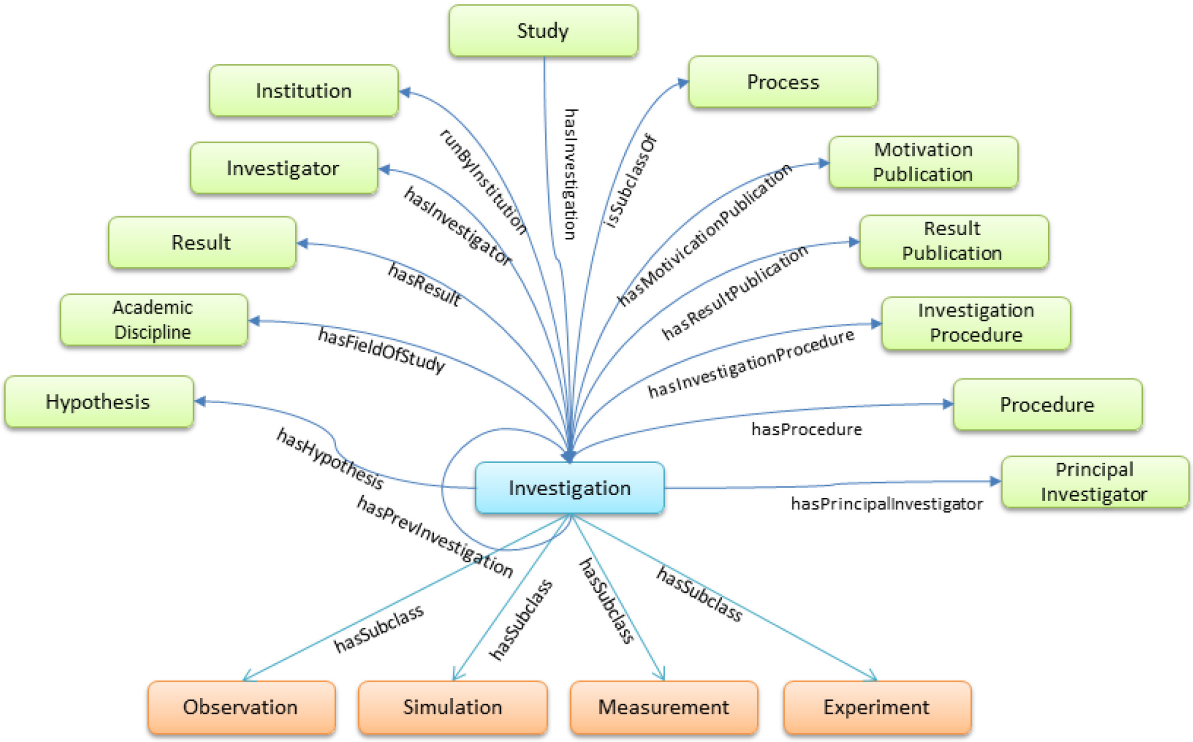


Figure 16: Graph representation of main relations of an Investigation

The investigation concept is also presented in SUMO. In the case of SUMO, the investigation is defined as the verb *Investigating*. For this reason, the cosi:Investigation is defined as sameAs sumo:Investigating, plus the additional properties. With these summaries, defining the investigation in Description Logic notation is as simple as the definition DL1:

$$\begin{aligned}
&\texttt{cosi:Investigation} \equiv \texttt{sumo:Investigating}\\
&\texttt{cosi:Investigation} \sqsubseteq \texttt{cosi:Process}\\
&\qquad\sqcap \geq_1 \texttt{cosi:hasInvestigator.cosi:PrincipalInvestigator} \quad \text{(DL1)}\\
&\qquad\qquad \sqcap \ni \texttt{cosi:hasProcedure.cosi:Procedure}\\
&\qquad\qquad \sqcap =_1 \texttt{cosi:hasHypothesis.cosi:Hypothesis}
\end{aligned}$$

The formalization in Description Logic notation provides the basis for the creation of the OWL notation. The following code is the representation in Turtle notation of the *cosi:Investigation* definition

```
### cosi:Investigation
cosi:Investigation rdf:type owl:Class ;
          rdfs:label "Investigation"@en ;
          owl:equivalentClass sumo:Investigating ;
          rdfs:subClassOf cosi:Process ,
              [ rdf:type owl:Restriction ;
                 owl:onProperty cosi:hasProcedure ;
                 owl:someValuesFrom cosi:Procedure
              ] ,
              [ rdf:type owl:Restriction ;
                 owl:onProperty cosi:hasInvestigator ;
                 owl:onClass cosi:PrincipalInvestigator ;
                 owl:minQualifiedCardinality
"1"^^xsd:nonNegativeInteger
              ] ,
              [ rdf:type owl:Restriction ;
                  owl:onProperty cosi:hasInvestigator ;
                  owl:someValuesFrom cosi:Investigator
              ] ,
              [ rdf:type owl:Restriction ;
                  owl:onProperty cosi:hasHypothesis ;
                  owl:cardinality "1"^^xsd:nonNegativeInteger
              ] ;

        prov#definition """Definition 2: Investigation is an
examination process in support of a hypothesis. It is part of a
systematic study and adheres to a scientific procedure
An investigation has a title, a description and is identifiable. It
has at least a principal investigator role and an undefined number of
additional investigators"""@en .
```

In this section we have discussed the *Investigation* class and in the scope of COSI formalization. To provide the OWL definition of this class, we discussed the following steps:

a) Define conceptual definition for the concept
b) Evidence relations to other concepts and denote the properties of the new class
c) Coin the definition in DL
d) Transform the DL definition to OWL 2 DL (expressed in Turtle Notation)

As we progress in this section, the full set of concepts of COSI will be explored. In the upcoming examples, the Description Logic notation and the OWL class definitions in Turtle notation are left outside of the text. The full OWL representation is documented in *Appendix C, Core Ontology of Scientific Investigation in Turtle Notation*.

The next section is dedicated a set of canonical classes that support the definition of other concepts in the COSI formalization. This set of classes represents a group of basic concepts. These concepts can be extended by other classes to inherit basic properties and allow categorization, a process that also influences ontology alignment.

4.2.2.2 Foundation Concepts

In the previous definition of *cosi:Investigation*, a reference was made to a parent class denoted as *Process*. The class name was underlined (see Definition 2) to show that the concept is still to be defined in the scope of this ontology.

SUMO provides a total of six instance definitions for the term *process* as a noun. In addition, seven more definitions for the same term exist as a verb. The closest definition to the desired reuse in the COSI context is the characterization of a process as *a sustained phenomenon or one marked by gradual changes*[101]. This definition relies on the concept of a *phenomenon,* which in turn is defined as a process in SUMO. Such a definition in Description Logic would lead to loop-hole.

A more technical definition for a *Process* is found in "Quality Management Systems-Fundamentals and Vocabulary" ISO 9000:2005 (International Organization for Standardization, 2005), which is ported in *Definition 3*.

> **Definition 3:** ***A process*** *is an activity or a set of activities that use resources to transfer input to output*

ISO 9000:2005 is also used in the definition of a *cosi:Procedure* as well.

> **Definition 4:** ***Procedure*** *is a specific way to carry out an activity or a process*

Definitions **3** and **4** rely on *activity* as a concept which leads to the next definition. The definition of procedure is necessary for further definitions documented in Section **4.2.2.4** *Investigation Subclasses and Procedures*.

In analysing the definition of a process, one could notice the relation to two very important concepts: *input* and *output*. Considering our digital domain of operation, both input and output may be represented by digital data objects. The information on the *input* data object and the *outcome* data is crucial to the contextual and provenance information related to an investigation. Inputs are data objects operated on by any process or activity. Outputs are defined as the data object resulting from a process or activity. These concepts can be mapped

[101] SUMO synset 100029677 - http://sigma-01.cim3.net:8080/sigma/WordNet.jsp?synset=100029677; Retrieved February 2015.

to the *Entity* definition of the PROV Ontology, which is a very broad concept. Definitions of *activity* and *entity* will be included from the PROV Ontology (Lebo, et al., 2013). See Figure 17 for an overview on the relations between Entity and Activity.

> **Definition 5:** *An* ***activity*** *is something that occurs over a period of time and acts upon or with entities*

> **Definition 6:** *An* ***entity*** *is a physical, digital, conceptual, or other kind of thing with some fixed aspects*

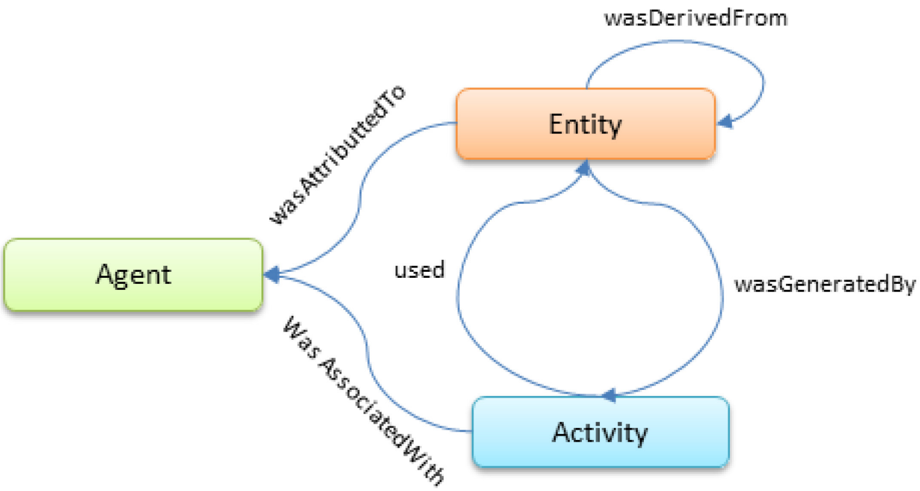


Figure 17: Provenance fundamental elements in COSI, Entity, Activity and Agent

In defining the class *cosi:Investigation*, we noticed that the class *Process* is defined as a super class of *Investigation*. It is marked same as a *prov:Activity* with the additional property *cosi:hasActivity* indicating that it may contain a set of activities (*cosi:has Activity* is defined in domain *Process* and range *Activity*). This axiom is a derivative of the Definition **3** where a process is considered as an activity or a set of activities. In addition, a process has an input data object and an outcome data object.

Table 11: Description of Process class in COSI

Class	cosi:Process
IRI	http://purl.org/net/cosi#Process
Equivalent to	prov:Activity
Definition	A process is an activity or a set of activities that use resources to transfer input to output.
Has Superclass	cosi:Onymous;

Has Subclass	cosi:Investigation;
Is in Domain of	cosi:hasActivity; cosi:hasInput; cosi:hasOutput

The process is listed as a child of an upper class referred to as *cosi:Onymous*. Onymous, as opposed to anonymous, is used in this context to denote an identifiable resource. The identification of a concept is important in the application and the portability of the model in a software implementation. Onymous is meant to be an upper class of resources that are expected to contain an identifier, a title (or an alternate title, e.g.: a name) and a description. Many classes in the scope of COSI extend this class to inherit these properties.

Definition 6: An **Onymous** is an entity that is identifiable, has a title and a description.

Table 12: Description of Onymous class in COSI

Class	cosi:Onymous
IRI	http://purl.org/net/cosi#Onymous
Definition	An Onymous is a *thing* that is identifiable, has a title and a description.

Has Superclass	Thing;
	cosi:hasIdentifier exactly 1 rdfs:Literal;

Described by properties	cosi:hasTitle, cosi:hasDescription;

Usually in OWL, all the concepts derive from a super class named *Thing*. With respect to the practical use, while defining different classes it can be noticed that similar concepts need similar identifiers. The *Onymous* class (see Table 12) is there to provide the basic identifying data properties. In a practical implementation, any class extending *Onymous* can be a *dataObject* which has an identifier and may have a title and a description. An alternative label for the title may be a “name”.

4.2.2.3 Investigator Roles

In the scope of COSI, the *Investigator* class is defined as a *role* that presents an active involvement in an investigation or a study. The definition of a *Role* is provided in the PROV Onotology. A *prov:Role* is the function of an entity or agent with respect to an activity. A role lives in the context of a usage, generation, invalidation, association, start, and end of an activity. The existence of this entity is also very important due to possible mapping of similar roles to a practical implementation in a software. In such case, roles should be mapped to an access control mechanism that allows and grants access to specific resources related to Investigations.

In COSI, an investigator is defined as:

Definition 7: *An **Investigator** is an entity actively involved in a research process. The investigator inherits all the properties of a Role class.*

Table 13: Description of Investigator class in COSI

`Class`	`cosi:Investigator`
`IRI`	`http://purl.org/net/cosi#Investigator`
`Definition`	`An Investigator is an entity involved in a research process.`
`Has Superclass`	`prov:Role;`
`Is in Range`	`hasInvestigator`

Information captured about the investigators enables valuable social contextual information related to a scientific investigation. The investigators are related to the *cosi:Investigation* class by the property *cosi:hasInvestigator*. This data property may link an investigator to a study. The class study will be described in in Section **4.2.2.6.**

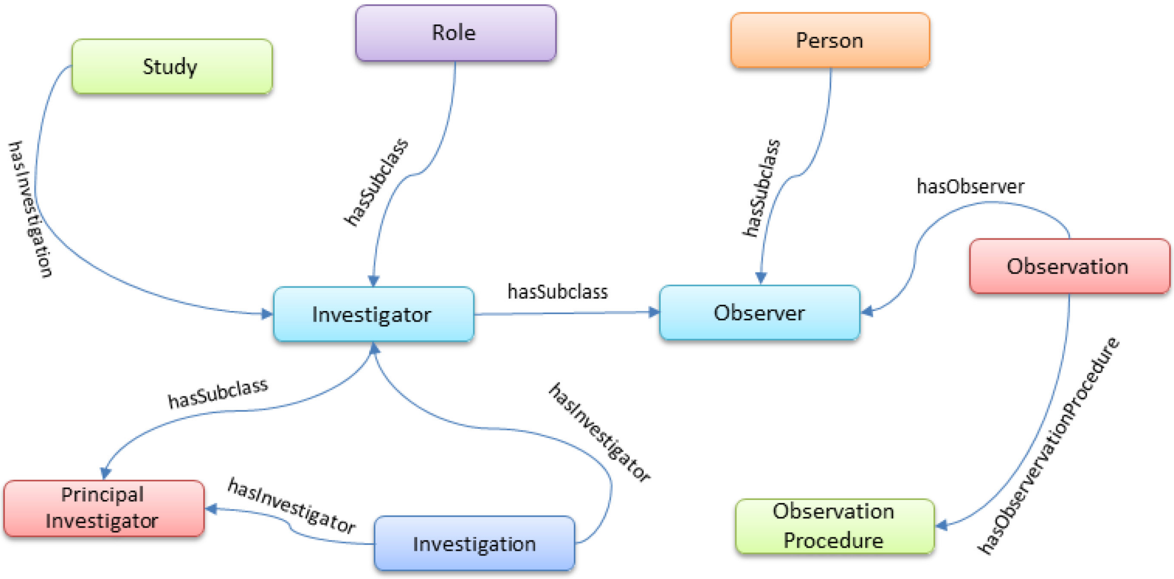


Figure 18: Investigator and Observer in COSI

A subclass of the *cosi:Investigator* is the *cosi:PrincipalInvestigator* defined as an investigator responsible for a specific investigation. The investigation entity relates different roles involved in an investigation. These roles are all grouped under the core role *investigator*. Sub classes of an investigator may include *students*, *assistants, researchers etc.* The organization of roles depends on the administrative organizations or investigation procedures. Since the COSI is a core ontology, additional roles can be defined by extending COSI. To show an example of how similar subclasses of *cosi:Investigation* are represented, the observer role is defined.

A *cosi:Observer* is a subclass of *prov:Role.* An observer is also related to a person therefore it extends the upper class *prov:Person* at the same time. The observer is a special role for an investigator that is involved in an *observation* process. Hence the definition of the *cosi:Observer*:

> **Definition 8:** *An **Observer** is a person involved in an observation. The observer becomes aware of through his senses.*

The observer is defined by the anonymous class *cosi:isObserverOf some Observation*. This property denotes that a *cosi:Observer* exists only if he/she has relation to an observation investigation (otherwise he may fit simply in the investigator role). The property *cosi:hasObserver* is the inverse property for *cosi:isObserverOf* and indicates an inverse relation of an observation and an observer. Inversion properties introduce a set of additional axioms in OWL 2 DL. The class which is referred from an object property in OWL is denoted as *the domain*. The class to which an object property relates, is denoted as *the range.* The inverse properties automatically populate the system with a set of axioms indicating that the invers property will also inherit the inverses of the domain and range of the original property. Therefore, if the property *cosi:hasObserver* has the domain *cosi:Observation* and range *cosi:Observer*, the inverse property *cosi:isObserverOf* will automatically have the domain *cosi:Observer* and the range *cosi:Observation* without the need to be denoted while defining the OWL class.

Table 14: Description of Observer class in COSI

`Class`	`cosi:Observer`
`IRI`	`http://purl.org/net/cosi#Observer`
`Definition`	`An Observer is a person involved in some Observation process that becomes aware of through the senses.`

`Has Superclass`	`cosi:Investigator;(prov:Role); prov:Person;`
	`isObserverOf some ObservationProcedure;`

`Is in Range`	`cosi:hasObserver`
`Is in Domain`	`cosi:isObserverOf`

4.2.2.4 Investigation Subclasses and Procedures

The SUMO upper ontology includes a broader set of *Investigating* subclasses. As the SUMO subclasses of Investigating are not necessarily related to a scientific investigation practice, (as SUMO is a foundation ontology), they are left outside from the COSI formalism. In some cases, *sumo:Investigating* is extended by investigations which may be found fit for some research activities, but are coupled to specific discipline usage. With this observation, the number of investigation subclasses can be extended depending on specific requirements of the investigation. Additional sub classes of Investigation can easily be added by extending COSI with new definitions. As the modelling effort is related to the development of a core ontology, the focus is on four principal investigation types that are: *Observation, Experiment, Measurement and Simulation*.

As presented in Figure 19, most of the properties can be generalized in the investigation concept. Differences between the subclasses of an investigation are related to the specific entities and procedures followed to run these investigations.

As discussed earlier, CSMD (Matthews, et al., 2010) and CSMO (Brahaj, et al., 2012) are two main models consulted in the formalization of COSI[102]. In these preceding models, subclasses of investigation differentiate from one another by means of their properties. This representation is shown in Figure 19 and presents the formalization of CSMO.

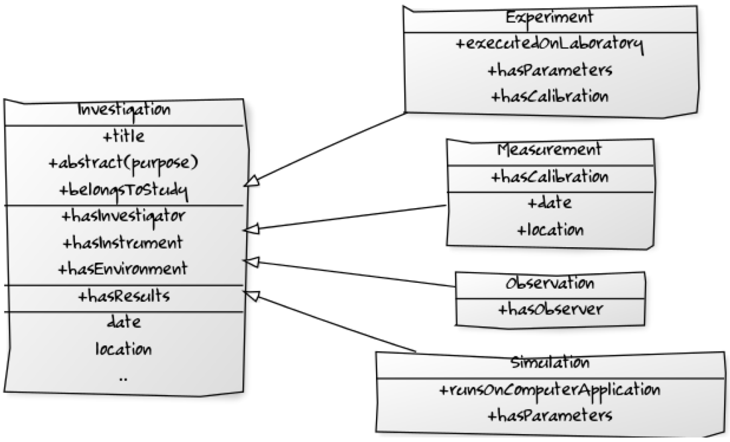


Figure 19: Investigation and Investigation Types based on the CSMO, the pre-formalization of COSI.

In COSI the definition of each investigation subclass is extended to rely not only on the internal properties, but also on specific procedures followed in investigations. In some cases, an investigation classification can be defined by the role of the investigator, for example, an *investigation* can be considered as an *observation* if it relies in a human observer solely. The outcome of the investigation in this case is an evaluation based on human perception. In other cases, the investigation is strictly defined on the procedure it follows. As

[102] As mentioned in the State of the Art discussion in Section 1.2, the Core Scientific Metadata Ontology (CSMO) (Brahaj, et al., 2012) was the first formalism attempted to provide an ontological representation of a scientific activity. It was based in Core of Scientific Metadata Model (CSMD) (Matthews, et al., 2010) and as such, it was an orthodox representation of CSMD. COSI precedes CSMO in the fact that it provides a formal definitions of the concepts, but also as it is amended with information harvested by in person interviews and research.

an example, an *experiment* is different from a *simulation* based on the way each of them is executed. Experiments rely in concrete controlled environments, usually a laboratory or an experiment facility. A simulation on the other hand, relies on a computing engine that is used to simulate scenarios and results. In defining similar sub classes of an investigation in COSI, an important criterion for the classification is also the *procedure* each of these investigations adheres to.

Observation is the first subclass of Investigation to be discussed. As it can be seen in Definition **9**, the classification of an observation is enforced by the presence of an observer and can adhere to a specific investigation procedure. In this case the procedure is denoted as an *Observation Procedure*.

> **Definition 9**: *An* ***Observation*** *is a form of investigation where results are assessed by a human perception adhering to an Observation Procedure.*

Table 15: Description of Observation class in COSI

Class	cosi:Observation
IRI	http://purl.org/net/cosi#Observation
Definition	An Observation is a form of investigation defined by an Observation Procedure and the perception of a human observer

Has Superclass	cosi:Investigation;(cosi:Process);
	cosi:hasObservationProcedure min 1 cosi:ObservationProcedure;

Is in Range	cosi:hasObserver
Is in Domain	cosi:isObserverOf

Measurement is another investigation procedure similar to the Observation. The major difference between these classes is the presence of instruments in measurements.

> **Definition 10**: *A* ***Measurement*** *is a form of investigation which makes use of instruments and is defined by an Investigation Procedure.*

Table 16: Description of Measurement class in COSI

Class	cosi:Measurement
IRI	http://purl.org/net/cosi#Measurement
Definition	A Measurement is a form of investigation which makes use of instruments and is defined by an Investigation Procedure.
Has Superclass	cosi:Investigation;(cosi:Process);
Is in Range	cosi:hasInvestigator

Beside the dependency to an instrument, a measurement is also related to time event properties. Typically, a measurement records states and conditions at a given point in time and the value would be the comparison of results in different intervals. The outcome data of a measurement is best used in analysis with other measurements data. The relations to the time and spatial properties are not expressed directly in the *cosi:Measurement* definition. These properties are inherited by the *prov:Activity* superclass, a super class of *cosi:Investigation* class.

In scope of COSI, an experiment is defined as:

> **Definition 10**: *An **Experiment** is a form of investigation executed in a research facility. The setup and results of an experiment are related to instruments and <u>Experiment Procedures</u>.*

An experiment is an investigation performed usually in a controlled environment. This might be a laboratory, or any other research facility. Experiments are conducted for the purpose of discovering or testing in order to assert or prove a premise. The process may be executed in repetitions with amendments of the parameters or instrument calibrations. The relation to an hypothesis is inherited by the *cosi:Investigation* class.

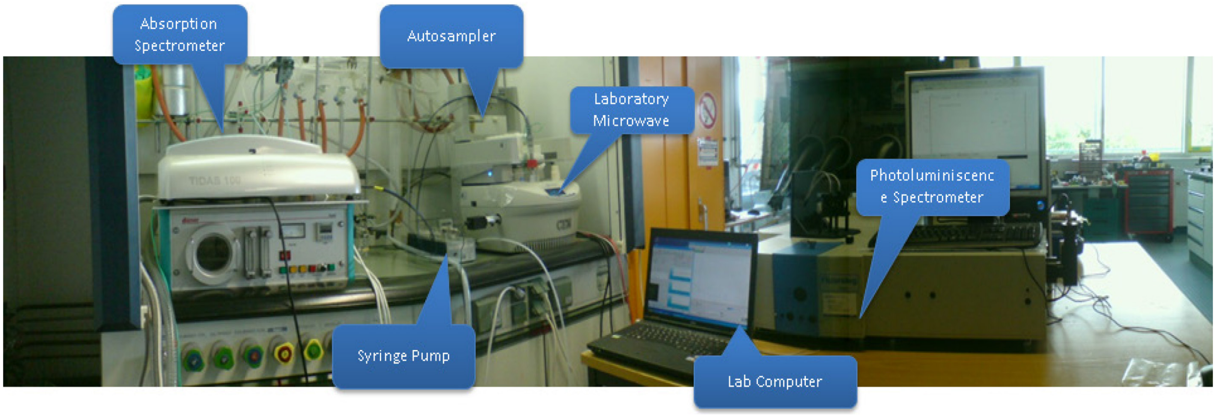


Figure 20: Laboratory environment comprising different instruments. Data from the experiments are stored in the laboratory computer.

Some other properties such as the relation to results, investigators or study are also inherited by the Investigation class. The property *cosi:hasParameter* is a new property and a reference to a specific set of Parameters which are used in a *cosi:Experiment* or a *cosi:ExperimentProcedure*. Iterations or relation to different procedures involved in an experiment are expressed through the experiment procedure.

Figure 20 shows a typical research setup in a laboratory. The final result generated from the experiment should contain information on each of the artefacts involved. This includes the instruments and also information on how these instruments are linked together and how was the flow defined.

Table 17: Description of Experiment class in COSI

Class	cosi:Experiment
IRI	http://purl.org/net/cosi#Experiment
Definition	An Experiment is a form of investigation executed in a specific a research facility. The configuration and results of an experiment are related to instruments and procedure used to run the experiment. The setup and the procedures of an experiment are defined by an Experiment Procedure.
Has Superclass	cosi:Investigation;(cosi:Process);
Is in Domain of	cosi:hasExperimentProcedure; cosi:hasResearchFacility; cosi:hasInstrument; cosi:hasParameter;
is in range of	isResearchFacilityOfExperiment

Instruments and the workflow followed while executing an experiment is very important. Outcomes of an experiment are correlated to the environment conditions and instrumental setup. Experiments are executed following a set of procedural rules that define the parameters, information or tasks to be passed from one participant to another for action. The instrumental set and other orders followed in executing the experiment are contained in the *Experiment Procedure.*

Definition 13: *An **Experiment Procedure** is a procedure or a set of ordered procedures involving artefacts of an experiment and their interoperable functionality*

As experiments are growing complex due to involvement of different instruments, the concept of scientific workflows has emerged to tackle the problem of excessive complexity. Workflow was defined in the business domain in 1996 by the Workflow Management Coalition (Workflow Management Coalition, 1993). It referred to the ability of running automated business processes based on well-defined services. The term workflow was borrowed by the research community and its main usage is documented in the Taverna Project (Wolstencroft, et al., 2013). Taverna's focus was the design of a methodology and tolls that support interconnection of experimental workflows across different Web Services. The result of the combination of these workflows would generate composite analysis pipelines. The main applicability of Taverna is related to bioinformatiocs. A typical Taverna workflow is composed of different ordered services that interact.In the scope of COSI, the *Experiment Procedure* is used to provide a broad generalization of the flow of an experiment. This is an equivalent of the workflow term used in Taverna. The artefacts in such a procedure are related to the concept of a Rig, or an ordered combination of instruments or services.

Table 18: Description of ExperimentProcedure class in COSI

Class	cosi:ExperimentProcedure
IRI	http://purl.org/net/cosi#ExperimentProcedure
Definition	An Experiment Procedure is a procedure or a set of ordered procedures involving artefacts of an experiment and their interoperable functionality.
Has Superclass	cosi:InvestigationProcedure;
Is in Domain of	cosi:hasOrderedProcedure; cosi:hasRig;
is in range of	hasExperimentProcedure

Definition 14: *A **Rig** is an ordered combination of artefacts participating in an experiment procedure*

Providing an organization of new instances as an ordered list is not a native concept in OWL. In the previous definitions, we have seen the utilization of foundation ontologies to inherit already defined concepts. In order to describe ordered lists as semantic graph, the COSI formalization makes use of Ordered List Ontology (OLO) (Abdallah, et al., 2010). OLO is based on two main classes; *olo:OrderedList* and *olo:Slot*. The *cosi:Rig* is implemented as an *olo:OrderedList* and it is basically defined as a set of instruments placed in ordered *olo:Slot*-s. The properties of *olo:index, olo:item, olo:length, olo:next, olo:ordered_list, olo:previous* and *olo:slot* allow iteration and extraction of the instances. *cosi:Rig* is not the only class defined as an extension of *olo:OrderedList*. The same functionality of providing an ordered list is also needed in the case of an experiment procedure. An example of a combination of the processes in an experiment procedure can be the definition of scientific workspaces similar to the definitions in the Taverna project, or any investigation which adheres to a set of iterations processes. Therefore, the *olo:OrderedList* is also used in the formalization of an Experiment Procedure as well (see Definition 13).

The last investigation subclass to be defined is the *Simulation*. A simulation as a form of investigation corresponds to a computer based technique of representing real scenarios by a computer program. A simulation imitates the internal processes and not merely the results of the scenario being simulated. Computer simulation packages will typically involve some simulation scenario with a set of initial parameters. The outcome of the simulation is a dataset representing the result of the simulation.

Definition 15: *A **Simulation** is a form of investigation created on a computer imitation environment representing real scenario based on digital modes*.

Simulation is an investigation executed through the imitation of the environment in a *computer application*. The environment is altered through *initial parameters* and the outcomes differ depending on alternations of the parameters, and what is more important, alternations of the digital model used in the simulation.

Table 19: Description of Simulation class in COSI

Class	cosi:Simulation
IRI	http://purl.org/net/cosi#Simulation
Definition	A Simulation is a form of investigation created on a computer imitation of a real scenario based on digital model.
Has Superclass	cosi:Investigation;
Is in Domain of	cosi:hasDigitalDomain; cosi:hasInstrument; cosi:runsOnComputer
is in range of	hasExperimentProcedure

4.2.2.5 Instrument

The *instrument* class is used to describe devices used in the course of an investigation. As mentioned previously, the information related to instruments used in the execution of investigations provides valuable technical information. The contextual information provided with the instrument should provide essential information on the instrument in case the same instrument is to be considered in reproducing the investigation. This information should include information on the vendor, model and other technical characteristics such as the throughput information, incoming and resulting information generated by the instrument.

Definition 15: *An **Instrument** is a device used in investigations*.

As defined in Definition 15, instruments are simple devices. They derive from the class *sumo:Device*. Depending on the nature of the instrument, additional properties can be used to better describe them. Table 20 lists the definition of class *cosi:Instrument* and the set of properties that can be used to annotate individuals of this class.

Table 20: Properties of an Instrument

Class	cosi:Instrument
IRI	http://purl.org/net/cosi#Instrument

Definition	An Instrument is a device used in investigations.
Has Superclass	sumo:Device;
Is in Domain of	cosi:belongsToResearchFacility; cosi:hasCalibration; cosi:hasModel; cosi:hasVendor; cosi:isPartOf;
is in range of	hasInstrument

To be noticed is that the definition for the instrument class is very loose. The set of properties describing an instrument is not restricted with additional new rules as we have seen with anonymous classes and existence quantifiers in previous examples. With this consideration, the set of properties for an investigation is to be extended based on the specific investigations' setup and contextual information that can improve the description of the instruments used.

From the set of properties presented in Table 20, of important relevance to result of an investigation is the property *hasCalibration*. Alternations on the calibration of the instrument will lead to different outcome results. The rest of the object properties for the instrument are relations to the research site (laboratory), and specific identifications such as vendor or model. An instrument can also be part of a *Rig.* This relation is presented by the property *isPartOf.* A *cosi:Rig* represents the combination of a set of instruments forming a platform or a pipeline, which can be used in an investigation.

The general class of instrument can be extended by more specialized instrument types such as digital instruments. Digital devices are based on electronic technology that generates, stores, and processes data in terms of two states: positive and non-positives. Instruments based on digital technology are based on computational operations which rely on operating systems. These groups of instruments, used in computer centric environment, inherit a set of additional properties which are presented in Table 21.

Definition 15: A **Digital Instrument** is a device based on a digital computational system.

Table 21: Digital Instrument properties

Class	cosi:DigitalInstrument
IRI	http://purl.org/net/cosi#DigitalInstrument
Definition	A Digital Instrument is a device based on a digital computational system.
Has Superclass	cosi:Instrument; cosi:hasOperatingSystemop exactly 1 sumo:OperatingSystem
Is in Domain of	cosi:dataOutput; cosi:hasMonitoredFolder; cosi:hasParameter; cosi:hasThroughput; cosi:returnsDataFormat

Digital instruments rely on the computational abilities. The definition of the class is also restricted to the presence of an operating system as well. Additional properties provide general information on the technical operation and capabilities of the instrument. The monitored folder is represented by a reference to a folder class[103], a literal or a URL which provides information on the data folder in a computer system. The monitored folder refers to the location where the input and output data are stored. This specific path can be synchronized with an ingest system to ingest data automatically in a repository, a process which is discussed in the Section 4.3 *Sheer Curation*.

4.2.2.6 Study

Each investigation is related to a *study,* a class that provides some general information on the scope of the running investigation. Each study can have one or more investigations. As we will see in Section 6.4, *COSI Portability to Other Scientific Disciplines*, a s*tudy* can also be related to the concept of a *project* in other ontologies[104] and is related to a concise focus of a research activity or knowledge acquisition.

In the scope of COSI, the Study is defined as:

Definition 16: A **Study** is a process focused on acquisition of knowledge.

Table 22: Description of Study class in COSI

Class	cosi:Study
IRI	http://purl.org/net/cosi#Study
Definition	A Study is a process focused on acquisition of knowledge.
Has Superclass	cosi:Process;
Is in Domain of	cosi:belongsToProgramme; cosi:hasInvestigation; cosi:hasInvestigator; cosi:hasLicenceInformation; cosi:hasRelatedMaterial; cosi:hasResult; cosi:hasStatus; cosi:hasTopic; cosi:runbyInstitution

The property *cosi:belongsToProgram* relates a study to a broader concept of research *programme.* The study is run by a *principle investigator* and has a specific lifetime. Properties related to the period of a study are not defined in the class per se, but inherited by *cosi:Process*.

[103] Folder Data Object and other filesystem referenced are discussed in the Section **4.2.8** Results and Data Objects

[104] See Section 6.4,*COSI Portability to Other Scientific Disciplines* for references for some mapping between ontologies

A *study* might be based and described by resources which are grouped as *related documents*. These materials can in turn be *publications, technical papers, study descriptions etc.* With respect to the interest in scientific investigation, additional valuable info of datasets released under a specific study is contained in the licencing information. This information provides general information on the access and reusability conditions[105] of the released research data. As the access control mechanism is tightly coupled with the infrastructure where the model is to be implemented, any property attached to the study would be merely descriptive. The implementations access control is part of a functional analysis which should be considered in the access control mechanism and is beyond the focus of this section. The access control's importance is stated in Requirement 10[106] and practical implementation of the access control are discussed in Sections 4.5 of this Chapter.

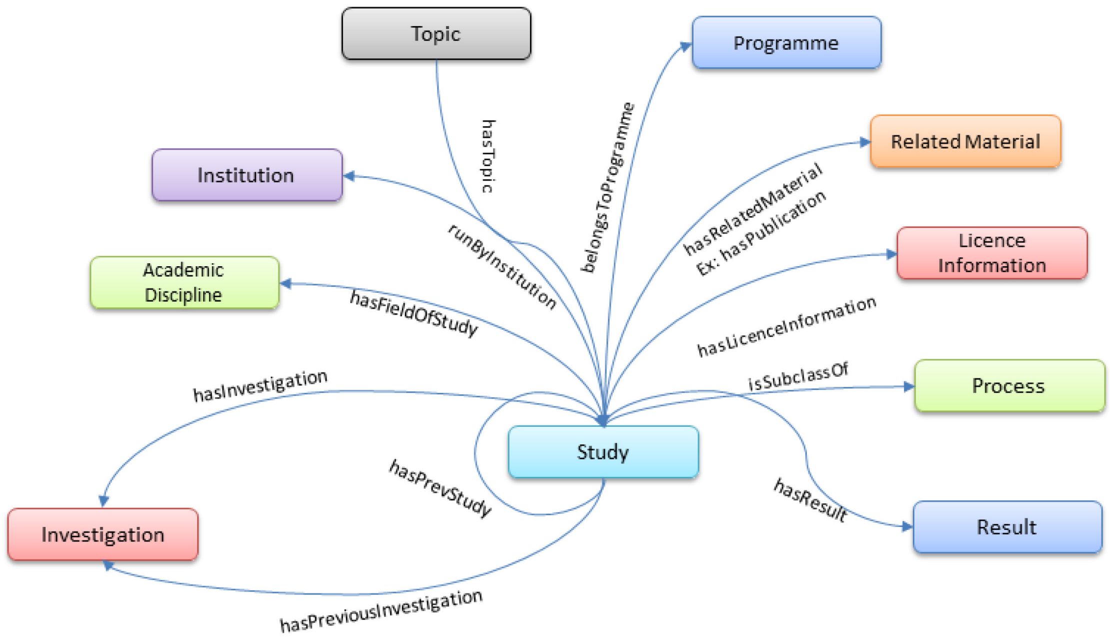


Figure 21: Study main relations in COSI

The study is also related to a specific institution, an entity that is inherited by *prov:Organization* and mapped to *cosi:Institution* class. A study is related to one or more institutions through the property *cosi:runbyInstitution*. In the same way, a study relates to investigations (as presented in Table 10: Description of Investigation class in COSI) and has a

[105] The Study entity might also contain additional administrative information such as information on the funding source, resources of the study etc. Since the work aims to provide a core model which can easily be extended, purely administrative properties of the study are left outside of the modelling analysis.

[106] See Section 3.1 Requirements and Annex A: Requirements

specific status. An important property with regard to the contextual information of result sets is the *cosi:hasTopic* which relates a study with a specific research discipline topic.

And the last property, *cosi:hasProgramme* relates a study to a *programme* concept, which is a container for a set of studies which might have a common subject or financed under the same theme. The programme entity in COSI will have basic literal data properties describing its theme, purpose and fields which allow for reference to funding or supporting organizations. In a similar fashion as in the case of *topic*, the *programme* can be extended by a vocabulary but this is beyond the concrete implementations desired in this work.

4.2.2.7 Topic

The *topic* object property provides information relevant to the particular investigation dataset. As the focus of the model is more centralized on the investigation and investigation datasets, the topic entity is designed in a minimalistic form. A minimalistic design of *topic* entities is not unusual in similar models, the same approach can be found in CCLRC Scientific Metadata Model (Sufi, et al., 2004), CERIF (Jörg, 2013) and other models.

The simplest representation of a topic would be a schema comprised of a set of entities pointing to a specific discipline. However, it makes more sense to link the topic entity with a standard vocabulary which may contain a comprehensive list of research topics. Similar vocabularies should be standardized and provide clear definition of each contained concept. As an example, in designing COSI the relation to topic is handled through the German standard classification described in the "*DFG Classification of Subject Area, Research Area and Scientific Discipline*" (DFG, 2008). Since the DFG classification is not represented in an ontological model, an ontological representation of this vocabulary has been also developed in the context of this research and is presented and documented in more details in Vocabulary of Scientific Disciplines (Brahaj, 2016). With the interest of this research, the alignment of the topic entity with the DFG classification of scientific disciplines or similar vocabularies is not part of the core ontology but the use is highly recommended.

4.2.2.8 Results and Data Objects

In the scope of COSI, a result is collection of data objects deriving from the execution of an investigation. The result consists of data collections and other additional eloquent information on the findings. These information address the initial hypothesis and in case of errors, the suspected cause for the errors.

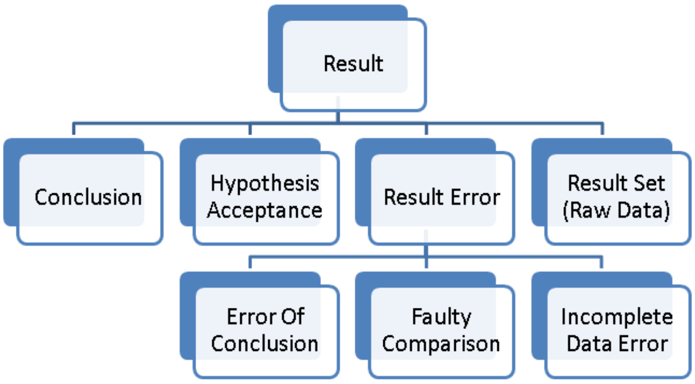


Figure 22: Result representation in COSI

Definition 18: *A **Result** is collection of data objects and conclusions deriving from the execution of an activity.*

The result is the representation of an investigation of activity, and therefore it does represent the activity with a conclusion. The conclusion itself is a reflection on the initial hypothesis and premises of the activity. As we can see in Table 23, the result provides a set of additional properties such as the link to a result error, in case the activity failed, link to a hypothesis acceptance and a reference to the full raw data generated in the course of the activity. The raw data or the Result Set is linked to the concept of *Data Object*.

Table 23: Description of Result class in COSI

Class	cosi:Result
IRI	http://purl.org/net/cosi#Result
Definition	A Result is collection of data objects and conclusions gathered as result of the execution of an activity
Is in Domain of	cosi:hasConclusion; cosi:hasResultError; cosi:hasHypothesisAcceptance; cosi:hasReport; cosi:derivesFromResult;
is in range of	cosi:hasResult;

The *data object* is based on the definition of NEPOMUK Information Element Ontology (NIE) (Mylka, et al., 2007). NIE is an attempt to provide unified vocabulary for describing native resources available on a computer environment. The data object in NIE is referred as *nfo:DataObject*[107] and *represents a native structure the user works with. The usage of the term 'native' is important. It means that a DataObject can be directly mapped to a data*

[107] NEPOMUK Information Element Ontology is referred to with the NIE acronym. It is composed of several documents, which together comprise the complete specification of the NEPOMUK Information Element Ontology Framework. In the definition of Data Object, the NFO prefix is used for NEPOMUK File Ontology.

structure maintained by a native application. This may be a file, a set of files or a part of a file. ... This class is not intended to be instantiated by itself. Use more specific subclasses.

As presented in Figure 23, the NIE-NFO includes a set of subclasses of the *nfo:DataObject* concept.

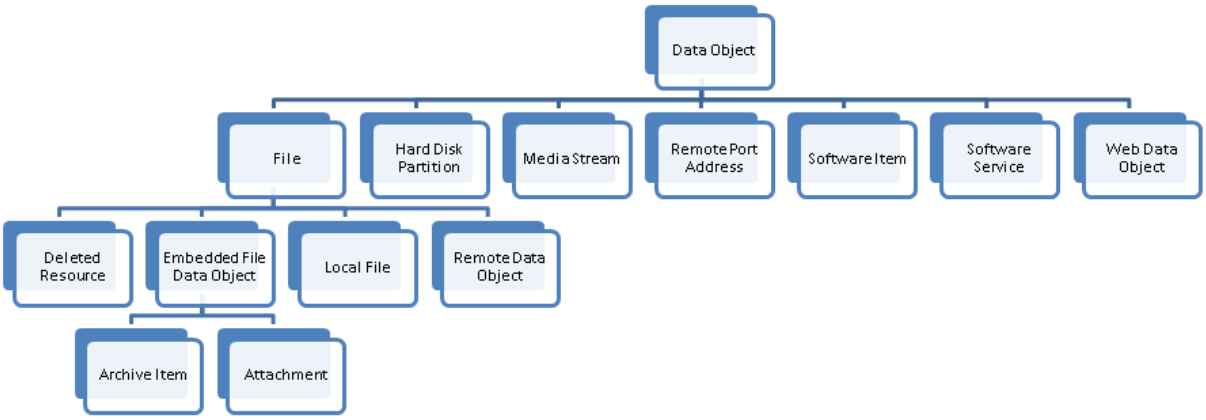


Figure 23: NEPOMUK Data Object hierarchy

In COSI, the raw digital object produced in the scope of an investigation, are related to a subclass of the *nfo:DataObject*. This might be a single *nfo:File*, a class that is defined as a finite sequences of bytes available from a durable storage medium. The *File* class includes subclasses of web documents and other resources resolvable via a URL, an integral feature in Linked Data. NIE allows an interpretation of the *nfo:File* to a group of *nfo:DataContainers* such as *nfo:Folder* or more specialized *nfo:Document*. That means, that by referencing the *cosi:ResultSet* to a *nfo:DataObject*, the implementation of the ontology in a practical use case will include any necessary storage reference. The data object reference is used to present a storage container for other properties in a scientific investigation. It might contain the datasets related to *calibrations*, *settings* and *parameters*.

In the context of COSI, access control mechanism and the relation to the data objects are to be implemented independently. Due to the important nature of the information stored within the data containers, versioning, persistent identifiers and other repository access information are designed and implemented as an extension of a repository feature. This theme is discussed in Section **4.4**, *Data Access*.

4.2.3 COSI Axiomatization

In the previous section, the fundamental classes of the COSI formalization are presented[108]. The formalization represents the inner modelling of the knowledge base related scientific investigation. The formalization allows referencing of provenance and contextual information related to a result-set. For example, knowing a data set formalized through COSI, one can query and derive information on the investigation it was produced, the researchers involved in the process, the institution supporting the investigation and more.

The ontology is defined through statements written with consideration to logic morphologic organization and obeys to strict Description Logic rules. The full set of OWL definitions is presented in the *Appendix C, Core Ontology of Scientific Investigation in Turtle Notation* together with the documentation of all the classes and properties. A visual presentation of all the classes and some relations is presented in Figure 24.

Altogether, the actual version of COSI has 120 classes. Its classes are based on concepts defined in the ontology and sometimes other classes extended from foundation ontologies. These classes are related to each other through more than 150 object properties. The classes also benefit from more than 90 data properties. Classes and their relations provide a total set of about 580 logical axioms. The implementation of the model as an OWL 2 DL ontology allows easy exploration, extraction of data and relationships about studies, investigators, investigations, publications, instruments, institutions attached to an investigation result.

The work on COSI is an attempt to simplify generation of valuable research data and enable their integration in the semantic web. Usage of semantic technologies improves access and allows a better specification of the contained data based on their meaning. In contrast to conventional separated instances of data sources, semantic web technologies can be used to implement a large knowledge graph of research data which may contain information from heterogeneous public repositories. COSI supports the organization of contextual research information alongside primary investigation's data. Once these data are published, any information system can query through protocols like SPARQL and retrieve more the necessary information on each entity. Moreover, semantic technologies allow for alignment of concept and searches in SPARQL can go merge different graph knowledge bases to retrieve additional similar or related information.

[108] For a full reference to the classes, please visit the online documentation website at http://purl.org/net/cosi

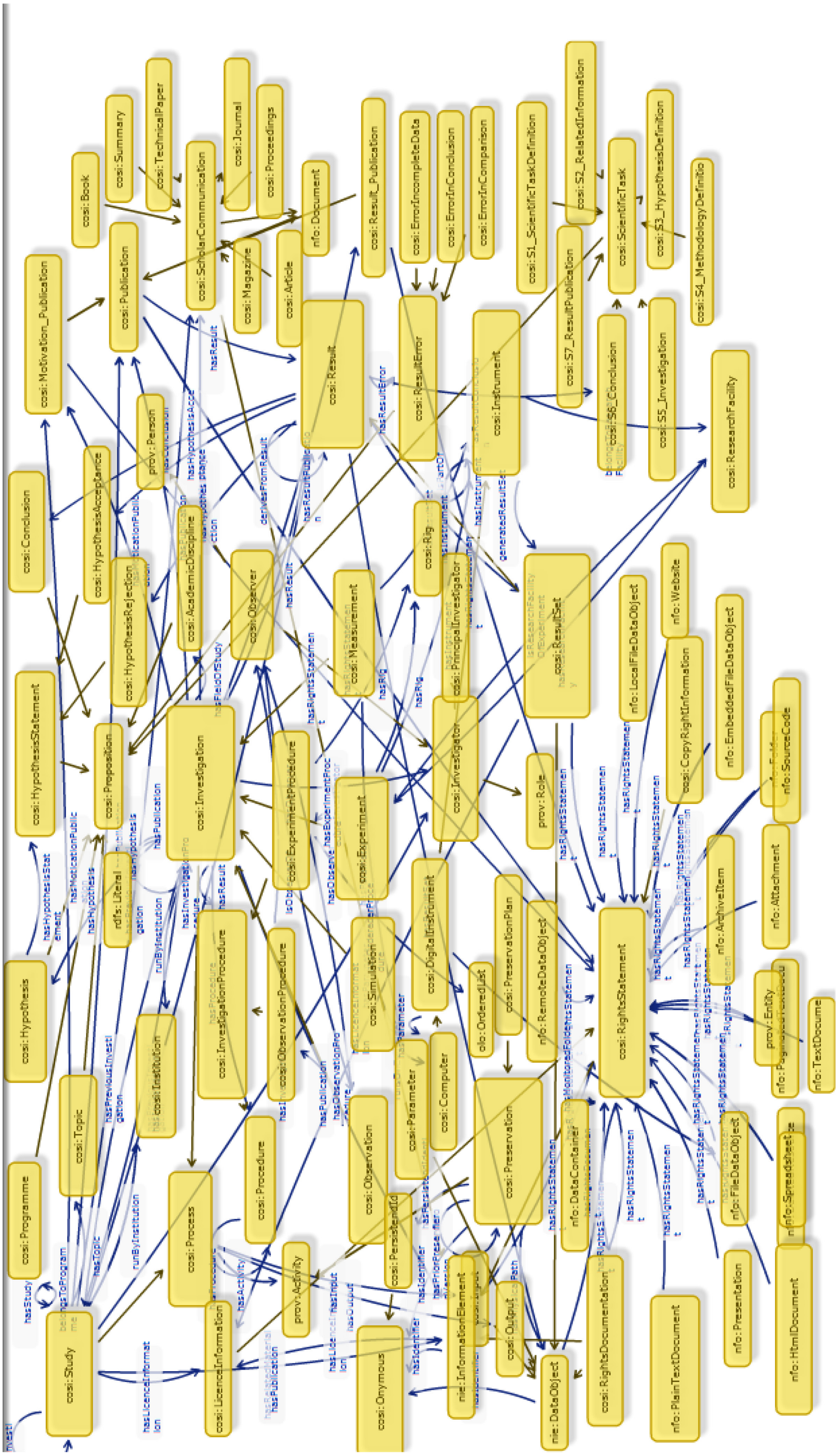
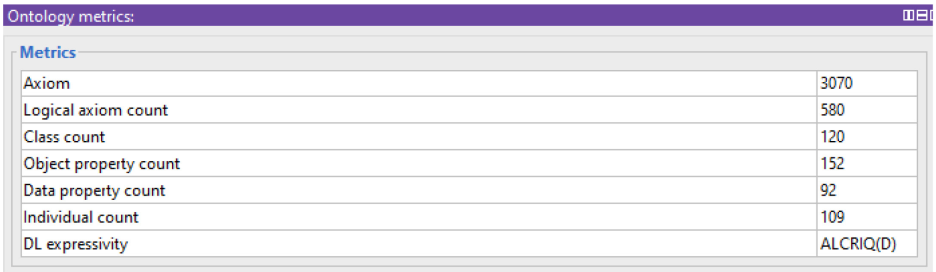


Figure 24: Visualization of the entities and relations in COSI



Ontology metrics:

Metrics

Axiom	3070
Logical axiom count	580
Class count	120
Object property count	152
Data property count	92
Individual count	109
DL expressivity	ALCRIQ(D)

Figure 25: COSI Ontology metrics as presented by Protégé

The next section is focused on the efforts needed to use the COSI in practice. The process is part of the digital curation and automation techniques.

4.3 Sheer Curation

This section deals with a practical implementation of COSI in the process of research data annotation. As we discussed in the previous section, COSI is composed of a broad set of concepts covering different aspects of an investigation process. The metadata provided in COSI contains references to administrative, descriptive, social and technical aspects related to the data generated from an investigation. An implementation of COSI allows the inclusion of a full set of metadata that can be used as "story teller" for an investigation. In order to provide this set of metadata, a feasible technique is needed that allows capturing all the activities reflected in the information model.

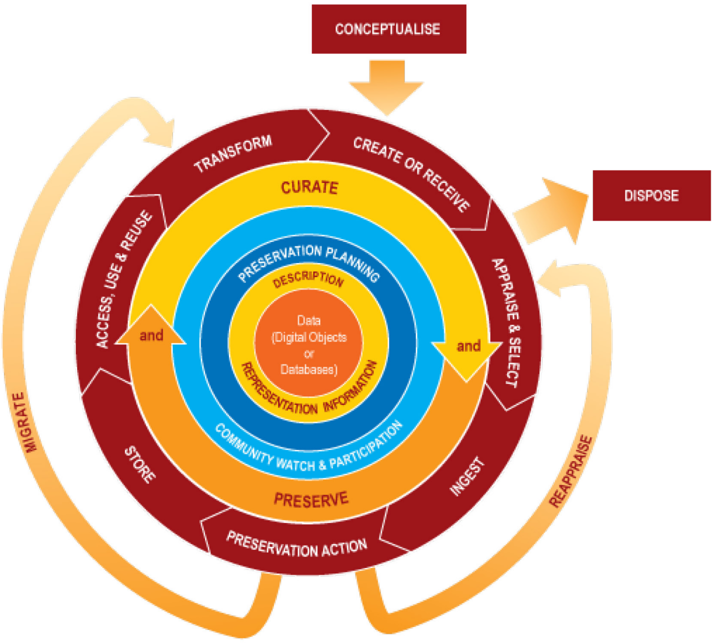


Figure 26: DCC Curation Lifecycle Model and stages evidenced. Modelling activities with COSI can be invoked at any stage, although it is beneficial to engage at the *Create or Receive* phase through Sheer Curation.

To capture and create this set of metadata, an annotation process that allows consumption of voluminous sets of metadata in a short time should be defined. As the number of investigations being run with the aid of computer supported environments is continuously increasing, a procedure that allows automation of metadata creation is highly desired (Gray, et al., 2005).

The process of the creating contextual, provenance and preservation metadata for digital assets is part of a discreet activity coined as *curation*. In the context of this work, the term *digital curation*[109] is used to reflect the process of creating the contextual, provenance and preservation metadata. The comprehensiveness of these metadata allows reuse of produced digital artefacts, may this occur in the present or the future. The digital curation activity has been subject of different normalization attempts. Among the most famous is the *Curation Lifecycle Model* (Higgins, 2012), a model published by the Digital Curation Centre (DCC)[110]. This model is a high-level representation of stages required for curation in the preservation process of digital data. The DCC curation lifecycle offers information on different stages of data preservation from initial conceptualisation through a whole possible iterative curation cycle.

The curation activities in the DCC curation lifecycle as presented in Figure 26 are denoted in the outmost layer. This set of activities starts with *Create or Receive*. Although all the activities represented in this layer introduce valuable provenance metadata, it is the first create-activity that can provide the fundamental contextual information on digital objects to be generated.

COSI can be used to reference provenance related to any of the activities documented in the DCC Curation Lifecycle Model, but its central metadata exist in the T1 moment, or the creation of the digital object. Through a semi-automation procedure, it is possible to enrich the digital data being created with contextual metadata at the point of creation. This is also a very rewarding procedure as researchers usually perceive metadata creation to be a time-consuming activity. Capturing the metadata at the time of creation of the digital objects is also an endorsement of the e-IRG Data Management Task Force who in their "Report on data

[109] There is a reflexive attempt to connect the term *curate (eng)* with *curare* (Lating and many Latin based languages). A curation process is not related to the *restore* process or *restaurate (lat*).
Curate derives from *cura (lat) – taking care; taking interest in something*.
Some other words rooted from *cura* are also *curios (eng) – taking in interest in something;* or *procure (eng) – to get by care*. With this consideration, *curare* is the act of administering or consuming a process attentively and with care.

[110] http://www.dcc.ac.uk/ - Digital Curation Centre (DCC) is a key recommendation by the JISC Continuing Access and Digital Preservation Strategy in the UK, which argued for the establishment of a national centre for solving challenges in digital curation that could not be tackled by any single institution or discipline.

management” note: *Researchers should be motivated to create metadata immediately and tool developers should add those descriptors that can be created automatically. “It is known that if metadata is not created immediately at resource creation time, the costs will increase rapidly and the quality decreases requiring costly curation efforts”.* (e-IRG Data Management Task Force, 2009)

Capturing the metadata at resource creation time requires an integration of curation activities in the normal investigation workflow. The activity of creating metadata in parallel with the process of creating and managing digital assets is known as *Sheer Curation*. The term *sheer* is a synonym for *transparently thin* or *pure*. In relation to curation activities, the term denotes the ability of a digital solution that is able to capture all the entities and activities encountered during an investigation process.

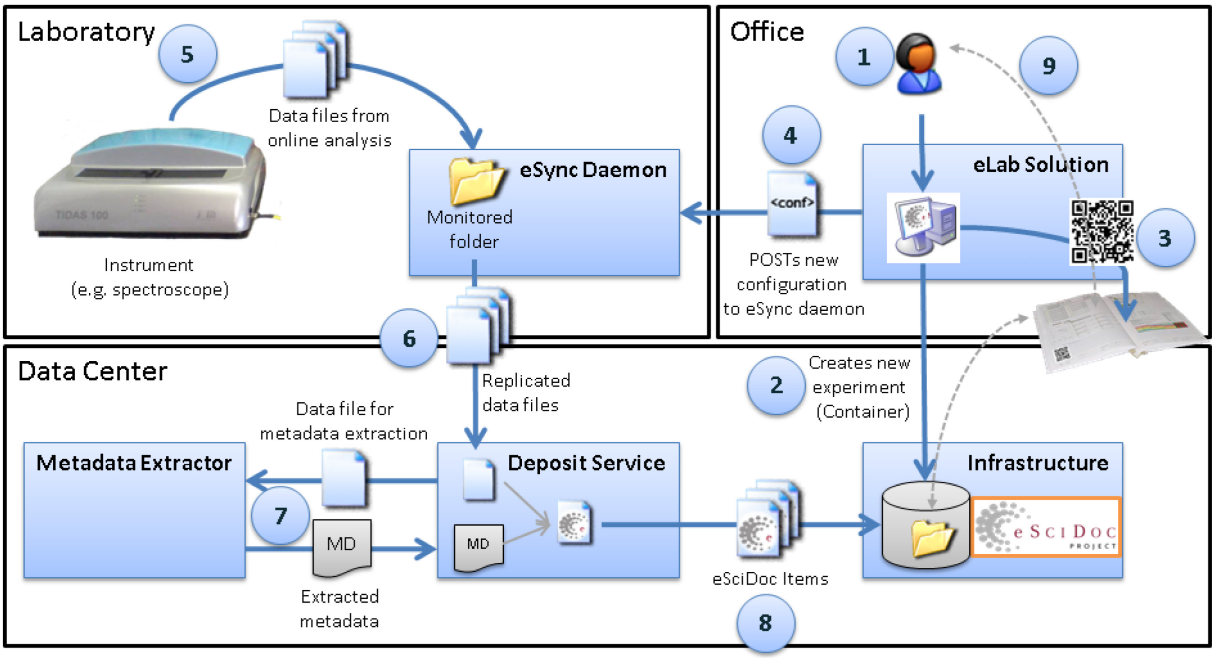


Figure 27: An Investigation Process and Sheer Curation through eSciDoc Infrastructure and COSI[111].

Figure 27 presents an integration of curation activities in an investigation workflow. The process is based on a laboratory environment backed by a data repository. The data repository is denoted by Infrastructure – eSciDoc, which is a research data management solution. In order to benefit from the RDF formalism and other benefits of the Semantic web, a triple store is attached to the eSciDoc repository. For the solution at hand, Sesame (Broekstra, et al., 2002) and a search layer, SIREn (Delbru, et al., 2010) are used. Sesame is desired to provide support for SPARQL capabilities to the data, although this implementation

[111] Credits for the image belong to Matthias Razum

provides some constraints in the access control to the digital objects[112]. The SIREn supports efficient indexing and querying of the RDF data guaranteeing a robust and scalable solution. The solution was also favoured since Lucene[113], core of the SIREn, is used in many open source implementations. The utilization of the triple store and the RDF search engine provides a fully semantic interface over the datasets stored in the repository.

In the course of this work, Sesame and SIREn are attached to the eSciDoc solution to provide another layer based on semantic technology. In fact, the repository where COSI may be applied can be any data repository with support for OWL/RDF, or the ability to be extended by modularized plugins. The trend of data repositories is shifting toward the RDF based repositories. Fedora Commons has already introduced the new organization of its internal data based with exclusive support of RDF. As Fedora Commons is a flagship data repository extended by many other data management solutions such as Islandora[114], eSciDoc[115], Hydra[116] etc, it is to be expected that the presence of RDF in digital repositories will rapidly increase in the short future.

Figure 27 represents an investigation executed with the support of a data repository and a virtual research environment tool referred to as eSciDoc Browser. eScidoc Browser is a data curation tool developed in the context of this work to support the integration and operation with an underlying repository. The activity visualized Figure 27 passes through the following steps:

1. In the first step, a researcher invokes the browser-based client and creates a new experiment within the context of a project or an investigation series
2. The experiment is represented by a folder in eSciDoc, which will later contain the captured data objects

[112] See Section 4.4 Data Access as well for some insight to the topic of data access in OWL and semantic technologies.

[113] https://lucene.apache.org/core/- Apache Lucene is a high-performance, full-featured text search engine library. It is a technology suitable for applications that requires full-text search, especially cross-platform.

[114] http://islandora.ca/ - Islandora is an open-source software framework designed to help institutions and organizations and their audiences collaboratively manage, and discover digital assets using a best-practices framework. Islandora was originally developed by the University of Prince Edward Island's Robertson Library.

[115] https://www.escidoc.org – Is a solution from FIZ Karlsruhe. eSciDoc version 1 was based on Fedora. The inprogress work of eSciDoc-Ng is based on a different solution, although by design eSciDoc can be coupled with Fedora Commons or any other data repository.

[116] http://projecthydra.org/ - Hydra is an ecosystem of components that lets institutions build and deploy robust and durable digital repositories. The project has originated as University of Alberta.

3. The experimental data is assigned an identifier so the experiment data may be referenced from a traditional paper-bound laboratory journal
4. The researcher then picks a predefined workflow of the investigation, or the order of a group of instruments (a so-called rig), which fits his needs.
5. As soon as the researcher starts the experiment, a synchronization process monitors one or more directories on the laboratory computer used by the instruments to store their measurements.
6. A specific process retrieves the data
7. Metadata are attached to the data based on the automatically captured contextual information like workflow, rig, instruments, users logged on to the system, investigation series, project, timestamp, etc., and by analysing the measured data.
8. Metadata record and the replicated data object are combined in a resource and stores the newly created item for preservation.
9. The researcher (or a colleague over the internet) may retrieve the data objects either by navigating through the data repository via projects and investigation series or by identifying the object in the laboratory journal

The process of creating and storing data in the repository needs to adhere to specific repository requirements. As described, the repository can be considered as a background service, providing storage, persistent identification, preservation and discovery of the content. These features influence the way digital objects are stored in a repository together with the additional related data sets contained within. The metadata, which in our case are represented in RDF structure, are a part of the digital object stored in the repository. Depending on the implementation, the digital object includes raw data, the metadata and additional functional info such as versioning, licensing, disposal information. Support for these services is provided in our scenario by the repository layer of eSciDoc.

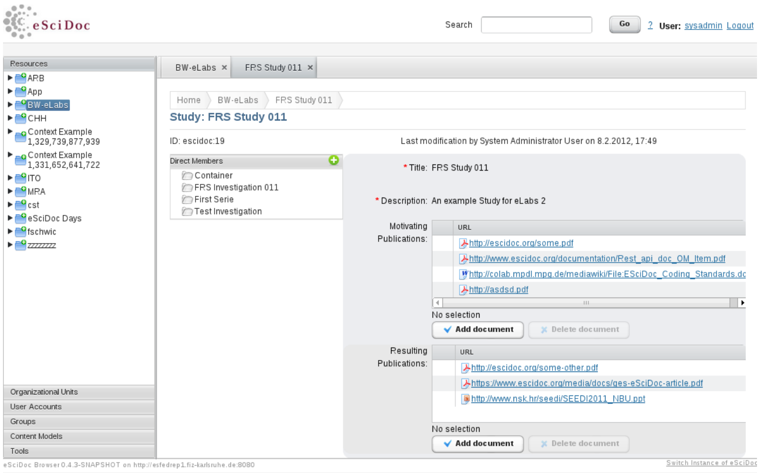
During the preparation and preservation of metadata related to an investigation, one or more semi-automated processes are incorporated in a pipeline that enables a sheer curation activity. In break down steps, the software components/services that enable this activity are three:

1) Interface to create and attach contextual information to the forthcoming investigation. The interface may also define provenance relations of the investigation to previous investigations or technical information such as the calibration of the devices, parameters used etc.
2) Retrieval of the generated raw data from the investigation environment
3) Technical metadata generation and ingest to the repository

These processes are to be supported by one or more software services. In implementation of a sheer curation activity in the scope of this work, the following software components were used: A metadata curation interface (eScidoc Browser), a *synchronisation service* and a *deposit service*.

eSciDoc Browser[117] as a metadata curation tool

Although, the process of extracting and storing metadata related to the running investigation is automated, there is always the need to provide a degree of intellectual curation (leading to a semi-automated approach). The first component in the sheer curation activity is related to the definition of a running investigation, and correlation of this investigation to a contextual set of metadata. A software interface would allow adding contextual information such as study under which the investigation runs, investigator in charge, publications it derives from, etc. By the end of the investigation, it is possible to define the access domain for the investigation data and attach information on the licencing and rights declarations. As a practical solution to improve the flow of a sheer curation tool, in the scope of BW-eLabs, a generic browser that allows the curation of related to investigations was developed. The generic browser was tied to the brand of the eSciDoc, a set of services that support a virtual environment for researchers. Figure 28 shows a visual representation of a study whose information is stored as RDF/XML.



Study: FRS Study 011

ID: escidoc:19

Last modification by System Administrator User on 8.2.2012, 17:49

* Title: FRS Study 011

* Description: An example Study for eLabs 2

Motivating Publications:

- http://escidoc.org/some.pdf
- http://www.escidoc.org/documentation/Rest_api_doc_OM_Item.pdf
- http://colab.mpdl.mpg.de/mediawiki/File:ESciDoc_Coding_Standards.dc
- http://asdsd.pdf

Resulting Publications:

- http://escidoc.org/some-other.pdf
- https://www.escidoc.org/media/docs/ges-eSciDoc-article.pdf
- http://www.nsk.hr/seedi/SEEDI2011_NBU.ppt

eSciDoc Browser 0.4.3-SNAPSHOT on http://esfedrep1.fiz-karlsruhe.de:8080

Figure 28: eSciDoc Browser visualizing results of a Study

Results and data generated during the runtime of an experiment are automatically added and visualized through the eSciDoc Browser. The same way a Study is represented, any other entity of COSI can be visualized with the eSciDoc browser. Data Properties and Object

[117] eSciDoc Browser is a Java based web application based on Vaadin. It was developed parallel to the research documented in this thesis to enable a Sheer Curation workflow. Source can be downloaded from https://github.com/escidoc/escidoc-browser (Accessed 2016)

properties define the descriptive fields for each of the forms. A curator or the investigator can easily add or modify whenever necessary the contextual metadata of a running or about to run investigation. The eSciDoc Browser is an ontology based metadata editor. The browser makes use of an ontology definition (COSI) and is able to edit instances and relationships of the ontology. Other operations such as semantic enrichments can be part of the intellectual curation in such a generic browser. Technically the generic browser uses an automatic technique to leverage relations of data knowledge base in a closed domain and at the same time, it can interconnect relations in an open infrastructure across different data repositories.

There are different alternatives to the use of eSciDoc Browser in a Sheer curation scenario. A very well-known tool among scholars is VIVO (Krafft, et al., 2010), a Cornell University initiative. VIVO supports recording, editing, searching, browsing and visualizing scholarly activity. Although VIVO per sè is focused on the conservation and documentation of academic personnel and academic publications (such as within an institution), core component of VIVO is a VITRO (Duraspace, 2015), a generic ontology editor. Through Vitro, one can create or load ontologies in OWL format, edit instances and relationships, build a public web site to display data and search the data (through Apache Solr). As Vitro is a generic editor, it can easily load the COSI ontology and allow creation and modification of instances based on the ontology.

Another potential alternative is Wisski[118], a Drupal based set of modules that may be used as a Virtual Research Environment (VRE) for managing scholarly data. Wisski is usually shipped with the Erlangen CIDOC CRM ontology (Goerz, et al., 2008), but it is possible to load any ontology formalisation. In Wisski, the ontology is seen as a set of primitive classes. Data entry forms are defined in a set of templates also known as *pathbuilders*. Each pathbuilder template defines a new data entry form for new objects created in Wisski. The solution can be easily adapted to provide a curation tool based on COSI or any other ontology model as well.

Synchronisation and Ingest services

In documenting the full set of data related to an investigation, beside the curation tool, a synchronisation and a deposit service are also used. Both services are related to the transfer of the investigation raw data from the investigation environment, e.g. a lab environment, to the data repository. The synchronisation service is as simple as the name. It basically retrieves the result data of an investigation from a specific network folder within the investigation setup to a folder accessible by the deposit service. Implementation of a synchronisation service can

[118] http://wiss-ki.eu/ Wisski is a German acronym for "Wissenschaftliche KommunikationsInfrastruktur", which can be translated as "Scientific Communication Infrastructure"

be as simple as an implementation of the rsync algorithm (Tridgell, et al., 1996) and there are different implementations already open and well maintained such as JarSync (Casey, 2013), Unison[119], DirSync Pro[120] etc.

The Deposit Service is in charge for the ingest activity to the underlying repository. In the scenario implemented presented in Figure 27, the deposit service is also in charge of a technical metadata enrichment process. The data retrieved from the investigation environment are analysed and technical information on the data per sè is extracted. In the implementations in the BW-eLabs, a service known as eSciDoc Ingest Client API[121] was used. This specific interface supports ingesting resources into the eSciDoc Infrastructure from different sources. It extends the eSciDoc Java Client Library, a simple Java application interface to eSciDoc Services, by adding a simplified approach for ingesting resources into eSciDoc. The API comes with two concrete implementations of an Ingester, a *DirectoryIngester* for ingesting directories and files from a local filesystem and a *ByNameIngester* that ingests resources of a given type from a given list of names. A more suitable solution for a Deposit Service would be any implementation based on the SWORD (Lewis, et al., 2012) protocol. SWORD is an acronym for Simple Web-service Offering Repository Deposit and is an initiative of the Joint Information Systems Committee UK (JISK)[122]. SWORD protocol defines an interoperable approach for deposit activities to repository platforms. Implementations based on the SWORD protocol, would support ingest of different digital object types in a consistent way and provide a unique protocol that once implemented in different repositories, makes it easier to provide similar generic sheer curation activities. At the moment, there are a number of implementations of the protocol for different programming languages such as Java, Python, PHP and support from different repositories such as DSpace, EPrints, DataBank, eSciDoc etc (Joint Information Systems Committee (UK), 2011).

The activity of sheer curation is an important concept, and an increase of the utilization of this activity in practice can highly impact the way investigations are executed and how knowledge out of these investigations is preserved. Practicing a sheer curation approach, influences quantity and possibilities of reusing research data, as researchers need less effort to create abundant metadata. The presented scenario is realistic and it is based on

[119] Unison File Synchronizer, University of Pennsylvania - Department of Computer and Information Science, Accessed October 2015, Available from http://www.cis.upenn.edu/~bcpierce/unison/lists.html

[120] DirSync Pro, Directory Synchronize Pro, Accessed October 2015, Available from http://www.dirsyncpro.org/

[121] "eSciDoc Ingest Client API" Accessed 2015, Available from https://github.com/escidoc/escidoc-deposit-api, Accessed October 2015

[122] Joint Information Systems Committee UK, Accessed 2015, Available from https://www.jisc.ac.uk/

different projects that share the same patterns in dealing with research results; they are all investigations executed in a digital environment and belong to the category of small-sciences as contrary to well-funded research laboratories with sophisticated Laboratory Information Management Systems (LIMS). The requirements and the research of this work were shaped from the following use case projects: *BW-eLab, eKinematix* and the attempted project *NanoCollect*. A brief description of these projects is offered in Section 5.8 *Evaluation in the context of an application and a task.*

As it can be seen, the sheer curation process is best supported by identifying existing practices of executing an investigation and mapping all the activities to stages that can be represented and fed as data in a repository. These stages provide references for activities and entities to be documented in the process of creating new digital data objects. The process of creating the new contextual information is run in parallel with the generation of raw digital data, and includes all the necessary information to reproduce the investigation. The metadata captured in this process is referenced to the digital object. The aim of sheer curation is to establish a solid foundation for other curation activities. These activities are shown in the DCC Curation Lifecycle model and related to alternations the digital assets may face in other processes. By providing the fundamental metadata and contextual information through sheer curation at the point of creation, further curation activities may be carried out by specialists at appropriate institutional and organisation levels, whilst causing minimum of interference to others. The combination of sheer curation and COSI as an information modelling is particularly beneficial in small sciences applications where the combination can be used as a laboratory information management system (LIMS).

4.4 Data Licensing and Rights Declarations

This section's focus is the incorporation of licencing information on any data sets generated through implementations of COSI. Covering issues on Data Licencing is a prerequisite stated in *Requirement 11: The approach shall provide information on licencing.* Data Licencing is an ever actual topic in data management. There is an increasing number of research funders requiring that data produced in the course of the research should be made available for other researchers. This data should be released with clear information on the terms of usage, mainly because data sets without explicit license are a potential legal liability when it comes to reuse or further elaborations. Research data should be released accompanied with information on usage possibilities, especially considering the fact that there is no default legal position on how research data may be used. (Some attempts have been made to harmonize the regulations related to research data, for example the *Berne Convention for the Protection of Literary and Artistic Works* (WIPO, 1979) provides a level of consistency among the countries that have signed it, but there are still exemptions defined in each national jurisdictions contributing to complexities and ambiguities surrounding the copyrights issues.) To avoid ambiguities surrounding terms of use of data, different licencing regulations have been created. The communication of the terms of use for each research data set is done by pointing to specific *licences*. A licence in this context is a *legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held* (Ball, 2014). Although commonly referred to as licencing information, the correct term for the information on the use of data is *rights declaration*.

The rights declarations convey information on rights held, waived or licensed. Normally the declaration of the licence is provided in the metadata of documents, inline in webpages or in noticeable sections in other forms of publishing (such as print). With regard to information modelled in OWL and represented through RDF serializations, the problem is a little bit trickier. Although the information expressed in RDF is expected to be publicly available, may this be through SPARQL queries, or complete dataset dumps, information on the licencing should be always present. Information extracted from SPARQL or within files is usually composed of granular RDF statements. As we saw in the previous chapters, these RDF statements are expressed as triples, a very simple statement containing a subject, property and object. Triples are used to describe resources, which are in turn identified by URIs. The resources might be data or objects. The triple statements describing them will be metadata if data are being described, or merely data information in case of objects. Attaching licence information to each triple would lead to massive overhead and to non-feasible solutions. There is no standard or best practice for a metadata convention pointing to a licencing information. Therefore, the licencing information may be provided in different organization levels of RDF.

In special cases, information on a resource or a group of resource result in an identifiable RDF Graph[123] or in an RDF dataset. In other more complicated cases, resources in RDF may be linked to the resources in other datasets, creating structures referred to as RDF mappings. In identifying licensable structures in RDF, Rodriguez-Doncel et al., (Rodriguez-Doncel, et al., 2013) evidence the following levels for consideration:

- Single RDF triples
- RDF graphs or RDF datasets, as collections of data
- RDF mappings as intellectual activity
- external resources referred by RDF

[123] The Resource Description Framework (RDF): Concepts and Abstract Syntax, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, defines an RDF graph as a set of RDF triples.

Within these groupings, a single RDF triple is usually not checked against protection of intellectual property on its own granular level, but rather evaluated at upper levels of RDF arrangements. Such arrangements are contained in RDF graphs or RDF datasets.

An RDF Graph or an RDF Dataset matches the intellectual property commonly referred as a *database*. The notion of a database in intellectual property is related to the creative and intellectual effort put in to create specific arrangements of data. This might be the aggregation of data on a specific resource, or filtering and cleaning of noise information from different sources with outcome a newly methodical dataset structure. The concept is categorized to be intellectual property and subject to rights declaration in many jurisdictions, including EU (EUR-Lex, 1996). RDF datasets combining data from different RDF graphs or datasets require the specific authorization from each source.

Collections of triples relating resources in two or more different RDF datasets create what is referred to as RDF Mappings. As this process is usually an intellectual process, it is to be expected that the organization per se falls under specific rights declarations. Considering the interoperability vision of Linked Data, RDF mappings are an important part of the Linked Data ecosystem.

The same interoperability envisioned by Linked Data is based on the ability to refer, or link through RDF properties to external resources. Creating resources dependent on external entities enables consumption of data from RDF sets under diverse terms of use, and the process should be done with careful considerations toward rights assertions.

In the scope of COSI, the RDF Mappings and external linking are not relevant (although should be taken in consideration by consumption of data modelled per COSI in specific implementations). Therefore, we discuss the provision of rights declarations only for RDF Triples and RDF Graphs/Datasets. Attaching licence information to RDF Data is not a trivial process. Although Linked Data is referred as the set of practices for publishing and connecting structured data on the web (Klyne, et al., 2006), there is no standard or a formal recommendation on attaching information on the rights of usage to data expressed in RDF. The community has been creative in recommending a number of solutions to the problem of attaching rights of information to RDF triples.

The simplest way of pointing a resource to a licence is through an additional RDF triple. Such a triple, does usually link through a property, denoting a rights declaration, to point to a specific licence type. The object of the triple, the reference to a specific licence type will usually be a URI reference to a licence description outside the existing dataset, and represented simply by a literal or a URI. The only way to identify if a resource has a licence, is not through the object of the triple, which might be any literal or URI, but through the property used. It is for this reason that this property should be in the best choice, a standardized property, or in absence of standardization, the most common property used for rights declarations. In an observation done by Rodriguez-Doncel et.al., (Rodriguez-Doncel, et al., 2013), to assess the RDF elements mostly used to specify a licence, it was found that the Dublin Core *:rights* element was the most prominent property. The results of this observation are shown in Table **24**.

The Dublin Core *:rights* elements is followed by the XHMTL *:licence* element and the Creative Common *:licence*, which is in fact an extension of the Dublin Core *:licence*. The observation was by assessing the presence of these elements in Sindice (Oren, et al., 2008), an RDF search engine that is no longer operational. Other licencing elements from many other vocabularies were evaluated but their presence was not of any significance and worth consideration (Rodriguez-Doncel, et al., 2013). Some of the properties checked were *premis:licenceTerms* from PREMIS, *doap:licence* from Description of a Project ontology etc..

Table 24: Relative use of licencing terms in Linked Data **(Rodriguez-Doncel, et al., 2013)**[124]

Vocabulary	**Element**	**Usage**	**Usage (%)**
Dublin Core	rights	5,905,519	59%
XHTML	licence	3,825,939	38%
Creative Commons	licence	263805	3%
Dublin Core	licence	32,922	neglectable

In the scope of COSI, the *dc:rights* property is used in practical implementations and recommended for use whenever a rights declaration is needed. The property is defined in the Dublin Core as *information about rights held in and over the resource*. The definition is broad and allows use with any specific types of applicable licences. It is extended by two additional properties, *dc:"Access Rights"* and *dc:licence*. Other vocabularies have relied on the *dc:rights* or *dc:licence* and extended their own properties. The Creative Common *cc:licence* property documented in Table 24 is a derivative of the *dc:licence*.

Defining a licence on a resource base leads to a very limited approach. Applied to knowledge bases with a large number of resource instances it will lead to a lot of overhead and redundancy. To challenge this problem, the concepts of RDF Graph, or RDF Dataset are used to attach a licence to larger arrangements of resources that fall and may share the same

[124] The usage in percentage column contains round up values. In a first observation, seems the usage exceeds 100%, but the table is documented as published by the authors.

licence. An RDF Graph can be obtained by querying instances that share similar attributes, while RDF Datasets may be represented by physical arrangements of the digital data, such as files, repositories etc. To provide valuable information on RDF Graphs and RDF Datasets, different ontology vocabularies such as Vocabulary of Interlinked Datasets (VoID) (W3C, 2011) are also presented in the Linked Data Community. VoID is a work in progress vocabulary for expressing metadata about RDF datasets. The organization of VoID is based on the definition of an RDF Dataset, which is slightly different, with a little bit more elaboration than the definition used so far in this section. A dataset in VoID *is a set of RDF triples that are published, maintained or aggregated by a single provider* (W3C, 2011). In VoID, a dataset is an instance of the *void:Dataset* class. As such, a *void:Dataset* instance is a single RDF resource, therefore attaching rights declaration to such a resource can be accomplished through the *dc:rights* property. The VoID vocabulary is referred by COSI formalisation, to allow curators of data to easily include licence information attached to specific datasets. The organization of these datasets is can be physically defined by splitting RDF files in desired triples, or in more elaborated cases, by deriving all the results and resource pertaining to a specific Study, Investigation or Result. For this reason, each of these entities can be mapped to licence information that can be inherited by dynamic datasets generated through SPARQL queries. The example below shows in Turtle notation how a *void:Dataset* resource can be attributed to a specific licence information.

```
      :Investigation a void:Dataset ;
          dc:rights     <http://www.opendatacommons.org/odc-public-domain-
      dedication-and-licence/>;
          dc:″Access Right″ """To the extent possible under law, The
      Example Organisation has waived all copyright and related or
      neighboring rights to The Investigation51 Dataset.""";
  .
```

The aforementioned example, points that the rights of use for the specific resource are described in a document found under a specific URL. Such document should provide information on what rights are held, waived or licenced. Neglecting the presence of closed linked data (that will be discussed in the Access Control Mechanism section), the most common and meaningful data licences to be used when releasing Linked Data are:

- Public Domain Licenses
- Attribution Licenses
- Share-alike Licenses

Public domain licences waive all the intellectual property and neighbouring rights such as database rights[125] of the dataset. The most well-known choices for public domain

[125] The notion of a database is related to the creative and intellectual effort put in to create specific arrangements of data. This might be the aggregation of data on a specific resource across different repositories, or filtering and

licences related to database and datasets are the Open Data Commons – Public Domain Dedication & Licence (ODC-PDDL)[126] and the Creative Commons CC0 Public Domain Dedication[127]. *Attribution licenses* also waive all the rights, but require a specific attribution. As examples we can mention any of the licences of Creative Commons beside CC0, the Open Data Commons Attribution License or commonly known as ODC-By[128] etc. *Share-alike licenses* also waive the rights, but require that derived works keep the license. Classic examples of share-alike licences related to databases and datasets are the Creative Commons Attribution-ShareAlike (CC BY-SA)[129], or Open Data Commons Open Database License (ODbL)[130]. From a technical perspective, modelling data through COSI enables any researcher to annotate the data with a rights declaration from the list of the aforementioned licences. In the XML ecosystem, there exist a set of more complex solutions that can be defined through Rights Expression Languages (REL). Attempts to port REL to RDF Schema have been successful and vocabularies such as Creative Commons Rights Expression Language (ccREL)[131] already exist. The integration of ccREL follows the same patter already discussed for an integration in COSI.

As conclusion, in this section the issue of data licencing and rights declaration is discussed. The topic is of crucial importance in the research data management, but unfortunately no standard recommendation exists for RDF implementations. In this section a feasible solution and an approach to apply tagging of RDF resources and RDF datasets with licencing information is discussed. Licensing information for linked data should be done at different granularity levels. A practical solution for the problem is presented and integrated in the logic of COSI where data can be annotated with a licence from the granular triples explaining a resource, up to more aggregated datasets deriving from a *Project, Study,*

cleaning of noise information from different sources and the provision of a newly ordered dataset structure. The concept is categorized to be intellectual property and subject to rights declaration in many jurisdictions, including EU.

[126] http://opendatacommons.org/licenses/pddl/1-0/ - The Open Data Commons – Public Domain Dedication & Licence is a document intended to allow anyone to freely share, modify, and use a specific dataset or database for any purpose and without any restrictions

[127] https://creativecommons.org/about/cc0 - CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

[128] http://opendatacommons.org/licenses/by/

[129] https://creativecommons.org/licenses/by-sa/2.5/

[130] http://opendatacommons.org/licenses/odbl/

[131] http://creativecommons.org/ns

Investigation or *a Result*. A list of typical licence definitions was also briefly provided with example how an implementation would look like.

4.5 Data Access

Although we are living a momentum of responsiveness toward open access and other open frontier movements, providing access control mechanisms to research data is still a hard requirement. Access control is usually compulsory due to institutional guidelines, policies that in most cases are driven by monetization needs, although other research related concerns exist. COSI inherit's the *dc:rights* property to address to a specific rights declaration. The rights declaration banner may contain information related to applicable licences, but is also is extended by a *dc:"Access Rights"* property. Since an ontology is basically a set of definitions, the reference of *dc:"Access Rights"* will not constitute a mechanism of data protection; in the best case, it will refer to information describing a specific access control mechanism at a resource level, or in groups of resources gathered in RDF Graphs. Despite missing access control mechanisms, measures that enable realistic access control need to be present in applications of semantic technologies as well.

From a software engineering perspective, the right layer to address issues with access control is the repository implementation. In such implementations, access control mechanism implemented in the repository will guarantee the correct data access operations based on a set of rules already defined or presented within the data. Such setup will work flawlessly in most data serialization formats and will definitely work with direct access to RDF serialization formats. Solutions such as eXtensible Access Control Markup Language (XACML) (Godik, et al., 2002) will provide declarative access control policy rules for RDF/XML serialization and a processing model describing how to evaluate access requests according to policy rules. A different outlook exists in case of information stored in other native semantic representations, where access control mechanism will be ignored by SPARQL queries and other ontology matching operations. A native access control mechanism is absent in the SPARQL domain and in more general in technologies used in the Semantic Web.

The problem is part of intensive research activity in the last years and many researchers have provided different recommended solutions to address the issue. Most of these recommendations are based on policy based solutions, to mention a few works *Social Semantic SPARQL Security for Access Control Ontology (S4AC)* (Villata, et al., 2011) or A *View Based Access Control Model for SPARQL* (Gabillon, et al., 2010). Other research has advocated toward encryption of RDF-Graphs, (see Partial Encryption of RDF-Graphs (Giereth, 2005)), to solve the access control problem.

Each of the aforementioned solutions can be used in applications of COSI in real word scenarios. In the use cases explored during this research, the access control mechanism has been embedded in the repository services (ex: eSciDoc) and the use of an additional search level (Siren) that adheres to access control mechanism incorporated in the repository solution. As there are cases where COSI may be applied directly and without the aforementioned setup (of eSciDoc services for example), additional practical implementations may be used. As a recommendation, can be the hidden declarative information provided through RDF *blank nodes*. One of the features of RDF is to express incomplete metadata through a construct known as *blank nodes*. Blank notes, provide statements on un-referenced resources. In these statements, we can point to the existence of specific attributes of a resource, without specifying or referring via a URL the resource itself. In fact, blank nodes are considered existential variables in the data. The use of blank nodes for data access control is a valid solution in cases where access to specific data needs be limited entirely. In these cases, the blank nodes help shield sensitive information organized in different RDF Graphs. For example, metadata on an investigation are presented, but without including internal information on the investigation entities (and results), through replacing the real investigation's identity with the blank node. Thus, someone without proper access may just retrieve general/public information about the investigation, but cannot know details and extensive information about the identity and inner findings. Blank nodes provide an easy implementation of protection for RDF Graphs and RDF Sets through internal organization of the data sets.

Chapter Summary

In this chapter, I presented COSI, a core ontology for the modelling of scientific investigations. COSI is a formalisation attempt of the finite investigation environment where research operations are handled. It is envisioned to allow the storage and representation of provenance and contextual metadata. This information is immensely valuable to the understanding of research results and it is a practice of good science. The ontological data model is formalized based on Description Logic axioms and expressed in OWL notation. By relying on Description Logic, the model allows the production of valuable output data where operations of reasoning and ontology alignment are native. This allows for the interoperability of the data across repositories and disciplines.

The application of the model is discussed and its benefits are best exploited by applying sheer-curation, a novel solution for semi-automated data gathering. A practical implementation of a sheer-curation activity is presented pointing to the benefits of the process and the simplicity of integrating it to virtual research environments. To show practicably of the model in real life scenarios, questions such as data access and data licencing are also briefly discussed by showing how COSI and the sheer curation activity can address these issues. Topics of ontology and concept alignment, advanced access methods and the future of repositories are discussed in Chapter **6**, Discussion.

“Should philosophy guide experiments,
or should experiments guide philosophy?”
Cixin Liu (Liu, 2014)

5. Ontology Evaluation

In the last chapter, the formalization of COSI and some of its main components were presented. In this chapter, COSI will be evaluated using an ontology evaluation methodology to measure the quality of the developed ontology and practical application. This evaluation is related to conformance to Semantic Web standards and to the Linked Data principles (Berners-Lee, 2006). To prove on the quality of the developed ontology, a set of good practices need to be followed. These good practices are to be analysed based on certain criteria to ensure that the developed ontology meets the standard’s and can be used in practical use cases as well as theoretical ones.

This chapter is dedicated to the selection of an evaluation methodology and the evaluation of COSI. The methodology for the evaluation will be discussed in Section **5.1**. Based on the selected methodology, a set of criteria and aspect are extracted. To evaluate each of these components, 23 methods defined in the evaluation methodology are analysed one by one. Section **5.1.3** presents the quantified metrics used in the evaluation and Sections from **5.2** – **5.7** explain the methodologies and the result of the evaluation. Section **5.8** briefly presents some applications were the ontology is applied to argue the assessment of the ontology in the context of an application and a task. The chapter ends with a summary of the evaluation in Section **5.9**.

5.1 Methodology

Methodologies on evaluation of ontologies are still a hot research topic. This is mainly due to continuous evolution of the Semantic Web standards. Nonetheless, some prominent evaluation methodologies are already distinguished in the semantic web community. From the set of methodologies, the most comprehensive research work is the *Ontology Evaluation* of Denny Vrandečić (Vrandečić, 2010). Vrandečić makes use of prior research on the topic to define initially four categories of ontology evaluation:

- *Ontologies can be evaluated by themselves*
  In this category, the Golden standard is considered. Golden standard is used in the sense of comparing the syntax in the ontology definition with the syntax specification of the formal language in which the ontology is written (e.g. RDF, OWL, etc.) (Brank, et al., 2005)
- *Ontologies can be evaluated in some context*
  The context is often defined based on the competency questions and artefacts used to develop the ontology
- *Ontologies can be evaluated within an application*
  Also known as application-based ontology evaluation (Brank, et al., 2005)
- *Ontologies can be evaluated in the context of an application and a task*
  This approach is also known as task-based ontology evaluation (Porzel, et al., 2005)

From the aforementioned categories, the evaluation of ontologies based on the framework proposed by Vrandečić (Vrandečić, 2010) takes in consideration only the two first categories. The proposed approach is based on the premise that each of the above categories gains from evaluating the previous category, i.e. every ontology evaluated within an application should have been evaluated by itself and with some context before that. Errors on the two upper most categories are discovered easier in contrast to a much more complex environment of an application or a task. Beyond the evaluation methodology endorsed of Vrandečić, Section **5.8** presents a set of applications where the ontology has been used with reciprocal projects covering different research disciplines. A brief discussion of the portability of the ontology to other science disciplines is offered in Section **6.4** *COSI Portability to Other Scientific Disciplines.*

In the effort to define a set of aspects for the evaluation of an ontology, Vrandečić analyses five prominent research literature. Each of these researches defined their own set of ontology quality criteria or principles for quality ontology assessment. The analysed literatures are (Gómez-Pérez, 2004), (Gruber, 1995), (Grüninger, et al., 1995), (Gangemi, et al., 2005), (Obrst, et al., 2007.). A list of the criteria defined under each of these publications is listed in Table 25[132].

Table 25: Ontology quality criteria in research literature

Criteria defined in (Gómez-Pérez, 2004)	**Criteria defined in (Gruber, 1995)**	**Criteria defined in (Gangemi, et al., 2005)**	**Criteria defined in (Obrst, et al., 2007.)**
• **Consistency:** capturing both the logical consistency (i.e. no contradictions can be inferred) and the consistency between the formal and the informal descriptions (i.e. the comments and the formal descriptions match) • **Completeness**: All the knowledge that is expected to be in the ontology is either explicitly stated or can be inferred from the ontology. • **Conciseness**: if the ontology is free of any unnecessary, useless, or redundant axioms. • **Expandability**: refers to the required effort to add new definitions without altering the already stated semantics. • **Sensitiveness**: relates to how small changes in an axiom alter the semantics of the ontology.	• **Clarity:** An ontology should effectively communicate the intended meaning of defined terms. Definitions should be objective. When a definition can be stated in logical axioms, it should be. Where possible, a definition is preferred over a description. All entities should be documented with natural language • **Coherence:** Inferred statements should be correct. At the least, the defining axioms should be logically consistent. Also, the natural language documentation should be coherent with the formal statements. • **Extendibility:** An ontology should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology monotonically. New terms can be introduced without the need to revise existing axioms. • **Minimal encoding bias:** An encoding bias results when representation choices are made purely for the convenience of notation or	• **Coverage** of a particular domain, and the richness, complexity, and granularity of that coverage • **Intelligibility** to human users and curators • **Validity** and soundness • evaluation against the specific use cases, scenarios, requirements, applications, and data sources the ontology was developed to address • **Consistency** • **Completeness** • the sort of **inferences** for which they can be used • **Adaptability** and reusability for wider purposes • **Mappability** to upper level or other ontologies	• **Cognitive ergonomics**: this principle prospects an ontology that can be easily understood, manipulated, and exploited • **Transparency** (explicitness of organizing principles): this principle prospects an ontology that can be analysed in detail, with a rich formalization of conceptual choices and motivations. • **Computational integrity and efficiency**: this principle prospects an ontology that can be successfully/easily processed by a reasoner (inference engine,classiffier, etc.). • **Meta-level integrity**: this principle prospects an ontology that respects certain ordering criteria that are assumed as quality indicators. • **Flexibility**(context-boundedness): this principle prospects an ontology that can be easily adapted to multiple views. • **Compliance to expertise**: this principle prospects an ontology that is compliant to one or more users. • **Compliance to procedures for extension, integration, adaptation**, etc. This principle prospects an ontology that can be easily understood and

[132] (Grüninger, et al., 1995) define a single criterion, competency that is also already included in the table of criteria in Table 25.

	implementation. Encoding bias should be minimized, because knowledge-sharing agents may be implemented with different libraries and representation styles. • **Minimal ontological commitment:** The ontology should specify the weakest theory (i.e. allowing the most models) and defining only those terms that are essential to the communication of knowledge consistent with that theory		manipulated for reuse and adaptation. • **Generic accessibility** (computational as well as commercial): this principle prospects an ontology that can be easily accessed for effective application. • **Organizational fitness**: this principle prospects an ontology that can be easily deployed within an organization, and that has a good coverage for that context

The criteria defined in different literature were summarized into a concise set. Eight criteria result from this literature survey*: Accuracy, Adaptability, Clarity, Completeness, Computational-efficiency, Conciseness, Consistency* and *Organizational fitness*. As it can be noticed, criteria given in the literature are subsumed by this set with some minor exclusion[133]. Therefore, a framework concerned with the quality of an ontology, should be based on an analysis if the above criteria are met.

5.1.1 Criteria for Ontology Evaluation

In this section, a brief description of each of the criteria evidenced by (Vrandečić, 2010) is presented. Under each criterion, there are several methods proposed that will be used in the evaluation of COSI.

- **Accuracy**
  Accuracy criterion conditions the conformity of the ontology to the knowledge of the stakeholders about the domain. The accuracy is connected directly to the correct description of ontology components which includes classes, properties, individuals, and derived axioms.
- **Adaptability**
  Adaptability criterion expresses the ability of the ontology in addressing different conditions within some anticipated tasks. This criterion is mainly related to the fact that ontologies are meant to be used on the Web, and their usage cannot be predicted.

[133] In fact, evaluation criteria that deal with the underlying language used for describing the ontology instead of evaluating the ontology itself are ignored. This due to the fact that most criteria were defined before OWL became widespread and other knowledge representation languages were actively used. Some criteria that are based on the ontology language, such as expressivity, decidability, complexity are disregarded (Vrandečić, 2010)

- **Clarity**
  Clarity criterion expresses the lucidity as to perception or understanding of the entities and properties defined in the ontology. The expressions defined in the ontology to name classes, properties and individuals should be unambiguous. This means that the definition of terms should be independent of the context and have interpretation by the users. Entities or complex axioms should be documented and this includes definitions, comments and labels in different languages as expected by the usage of the ontology.
- **Completeness**
  Completeness as a criterion, expresses the extended coverage of the ontology over a domain of interest. Completeness covers different aspects, such as completeness with regard to the language, completeness of key concepts representing the domain, completeness with regard to an application etc. Completeness is also a metric related to the granularity and richness of the ontology.
- **Computational efficiency**
  Computational efficiency is related to practicability of the ontology in real use cases. This is related to the ability of the tools to work with the ontology in particular the speed that reasoners need to fulfil the required tasks, be it query answering, classification, or consistency checking. The computational efficiency is important with regard to the complexity that will be inherited to the reasoner and affect on the knowledge processing in case of large graphs of information.
- **Conciseness**
  Conciseness evaluates whether the ontology includes irrelevant classes or properties to the domain of interest. The ontology should impose a minimal ontological commitment for better performance and simplicity.
- **Consistency**
  Consistency as a criterion expresses uniformity among the defined axioms. A consistent ontology does not permit for any contradiction. In contrast with accuracy, the consistency states that the ontology itself can be interpreted, while accuracy states the compliance of the ontology with an external source. Generally, consistency includes logical consistency and coherence.
- **Organizational fitness**
  This criterion is related to several measures that decide how easily an ontology can be deployed within an organization. This includes different pieces such as people (adaptation), tools (technical or economic decisions), technology and familiarity with the technology used in ontologies.

Checking if the fulfilment of the criteria is true for an ontology is not a simple operation. To assist in the process, (Vrandečić, 2010) breaks down the activity in *evaluation methods*. An evaluation method will assess a specific feature of an ontology or make it explicit. As presented in Table 26, each criterion has a number of methods. A method may be related to one or more criteria. The result of a method will provide an indicator for how well one or more criteria are met.

Table 26: Methods proposed to address each of the criteria for ontology evaluation

Ontology Evaluation Criteria	Methods proposed for the evaluation
Accuracy	Method 3: Look up names Method 13: Search for anti-patterns Method 14: OntoClean meta-property check Method 18: Explicit terminology ratio Method 19: Checking competency questions against results Method 20: Checking competency questions with constraints Method 21: Unit testing with test ontologies Method 22: Increasing expressivity Method 23: Inconsistency checks with rules
Adaptability	Method 6: Check name declarations Method 10: Check for superfluous blank nodes Method 13: Searching for Anti-Patterns Method 15: Ensuring a stable class hierarchy Method 17: Explicitness of the subsumption hierarchy Method 19: Checking competency questions against results Method 21: Unit testing with test ontologies Method 22: Increasing expressivity Method 23: Inconsistency checks with rules
Clarity	Method 1: Check used protocols Method 2: Check response codes Method 3: Look up names Method 4: Check naming conventions Method 6: Check name declarations Method 7: Check literals and data types Method 8: Check language tags Method 9: Check labels and comments Method 14: OntoClean meta-property check Method 18: Explicit terminology ratio
Completeness	Method 3: Look up names Method 6: Check name declarations Method 7: Check literals and data types Method 9: Check labels and comments Method 10: Check for superfluous blank nodes Method 11: Validating against an XML schema Method 12: Ontology complexity Method 15: Ensuring a stable class hierarchy Method 16: Measuring language completeness Method 17: Explicitness of the subsumption hierarchy Method 19: Checking competency questions against results
Computational efficiency	Method 6: Check name declarations Method 7: Check literals and data types Method 10: Check for superfluous blank nodes Method 12: Ontology complexity Method 16: Measuring language completeness

Conciseness	Method 5: Metrics of ontology reuse Method 10: Check for superfluous blank nodes Method 15: Ensuring a stable class hierarchy Method 17: Explicitness of the subsumption hierarchy Method 18: Explicit terminology ratio Method 20: Checking competency questions with constraints
Consistency	Method 3: Look up names Method 4: Check naming conventions Method 5: Metrics of ontology reuse Method 9: Check labels and comments Method 12: Ontology complexity Method 13: Searching for Anti-Patterns Method 14: OntoClean meta-property check Method 16: Measuring language completeness Method 21: Unit testing with test ontologies Method 22: Increasing expressivity Method 23: Inconsistency checks with rules
Organizational fitness	Method 1: Check used protocols Method 2: Check response codes Method 3: Look up names Method 4: Check naming conventions Method 5: Metrics of ontology reuse Method 8: Check language tags Method 9: Check labels and comments Method 11: Validating against an XML schema Method 19: Checking competency questions against results

5.1.2 Aspects for Ontology Evaluation

Mapping the evaluation methods with the criteria for ontology evaluation can be slightly confusing, especially considering that a method can serve as an indicator for multiple criteria. A better organization of methods for evaluation of an ontology can be arranged if methods are linked with *aspects*. In ontology building, each aspect describes some choices that are made during the design of the ontology. Evaluating different ontology aspects, allows evaluators integrate different evaluation results in order to achieve an aggregated, qualitative ontology evaluation. Vrandečić proposes six aspects in his framework for ontology evaluation: *Vocabulary, Syntax, Structure, Semantics, Representation* and *Context*.

- **Vocabulary**
  The vocabulary aspect is related to the set of names used in the ontology. This aspect covers concepts such as URI references, literals, datatypes etc. Evaluation based on this aspect for COSI is discussed in Vocabulary Quality, Section 5.2.
- **Syntax**
  The syntax aspect is related to the encoding language used to express the ontology. Evaluation based on this aspect for COSI is discussed in Syntactic Quality, Section 5.3.
- **Structure**

The structure aspect is related to the arrangement of the ontology graph in COSI. The structure can vary highly even describing semantically the same ontology. Evaluation based on this aspect for COSI is discussed in Structural Quality, Section 5.4.

- **Semantics**
  The semantics aspect is related to the formal meaning being represented by the ontology. Evaluation based on this aspect for COSI is discussed in Semantic Quality, Section 5.5.
- **Representation**
  The representation aspect denotes the relation between the structure and the semantics. Evaluation based on this aspect for COSI is discussed in Representation quality, Section 5.6.
- **Context**
  The context aspect is related to the features the ontology carries and a check against artefacts in its domain of usage. Evaluation based on this aspect for COSI is discussed in Contextual Quality, Section 5.7.

5.1.3 Quantification Metrics for the Evaluation Findings

Next section is dedicated to an evaluation of the COSI related to the aforementioned aspects. The applicable methods are presented with their definition, a brief description and the evaluation result. In order to quantify the COSI evaluation and provide conclusive remarks about the results, the following four metrics are used.

- **Verified**
  Indicates that the method is applied and the evaluation results are positive. The verified metric is used to indicate that the method result is found to confirm that no problem was found.
- **Inapplicable**
  Indicates that the method could not be applied for the ontology. This may be due to superseded requirements expressed in a method, or reasoning capabilities of the underlying language used in the ontology modelling.
- **Deferred**
  Indicates that the method is applicable but could not be verified. This may be related to technical or time constraints. Deferred results are not an indicator of a positive or negative response, it defaults to a future task.
- **Failed**
  Indicates that the method is applied and the evaluation results are negative. The verified metric is used to indicate that the method result is found to confirm that problems were found.

The quantification metrics will be used to provide an evaluation report based on each method.

5.2 Vocabulary Quality

Evaluation of the quality of the vocabulary aspect is associated to the terminology and naming conventions used in an ontology. This section is dedicated to a set of methods used for the evaluation of the vocabulary in an ontology.

5.2.1.1 Method 1: Check used protocols

This method is used to check the web protocol used in the ontology. Web browsers and servers use TCP/IP protocols to connect to the Internet. Some common web protocols are HTTP, FTP, POP etc: A premise of the Linked Data is the utilization of URI references to identify anything, from a person over an abstract idea to a simple information resource on the Web. Classes, property definitions and individuals of ontologies are also identified through URIs. The Web makes use of the URI as a global identification system. The global scope of URIs promotes large-scale "network effects" (W3C Working Group, 2014).

URI references are strings that start with a protocol. A classic example of a protocol is the HyperText Transfer Protocol (HTTP). If the protocol is known and applied by the ontology based application, then the application may resolve the URI, or in simple terms, use the URI according to the protocol to locate a document that contains more information on the identified resource. Hence the first method of the vocabulary aspect, is related to the protocol used in the internal URIs of an ontology. The definition of the method is as follows:

> **Method 1 (Check used protocols)**
> All URIs in the ontology are checked to be well-formed URIs. The evaluator has to choose a set of allowed protocols for the evaluation task. The usage of any protocol other than HTTP should be explained. All URIs in the ontologies have to use one of the allowed protocols.

Evaluation

COSI is based on the HTTP protocol and thus all the URIs are resolvable. All the URI references in COSI are based on HTTP. An example of a entity reference is: `http://purl.org/net/cosi#Study`

Verification of the well-form URI was done by checking a list of URIs used in COSI through the Apache Commons (Apache Foundation, 2016) *UrlValidator* class

Example:

```
UrlValidator urlValidator = new UrlValidator();
urlValidator.isValid("http://purl.org/net/cosi#Study");
```

Conclusion: *Verified*

5.2.1.2 Method 2: Check response codes

For every request made to a HTTP URI, there is a specific code returned by the server. This code is a machine processable message indicating the result of the call. Different codes indicate different operations. A HTTP 200 response code for example, will indicate that the resource was located successfully and a response containing the resource is being returned by the server. Similar to response code 200, there are a predefined set of codes with special meanings.

The second method proposed by (Vrandečić, 2010) is related to a set of expected response codes that indicate a successful URI retrieval. These response codes are connected to two types of resources on the Web; *information resources and non-information resources* (Bizer, et al., 2007).

Method 2 (Check response codes)

For all HTTP URIs, make a HEAD call (or GET call) on them. The response code should be 200 OK or 303 *See Other*. Names with the same slash namespace should return the same response codes, otherwise this indicates an error.

- Information resources.
  Normally, a URI identifying an information resource is invoked to obtain a copy or representation of the resource it identifies. The operations are also called URI dereference. In usual cases, the server generates the expected representation, a new snapshot of the information resource's current state, and sends it back to the client using the HTTP response code 200 OK.
- Non-information resources cannot be dereferenced directly. In these cases, a trick is used in practice to enable URIs identifying non-information resources to be dereferenced. Unable to send the representation of the resource, the server sends the client the URI of an information resource which describes the non-information resource using the HTTP response code *303 See Other*. This is called a 303 redirect. The redirect is leads to the information resource located in a HTTP response code 200 OK.

In accessing information on the entities and properties of an ontology, the expected HTTP response codes are *200 OK* or *303 See Other*. The second method is related to the checks of response code of the HTTP request.

Evaluation

The server that hosts COSI is configured to reply with a HTTP Response 200 OK (for the information resource) and HTTP Response code 303 See Other (for non-information resources). The ontology makes use of a Persistent URL under purl.net. As expected, the response code is a *302 Moved Temporarily* (from the persistent URI handler), followed by the expected *200 OK* or a *303 See Other*.

Conclusion: *Verified*

5.2.1.3 Method 3: Look up Names

In computer hypertext, a fragment identifier is a short string of characters that refers to a resource that is subordinate to another, primary resource. The primary resource is identified by a Uniform Resource Identifier (URI), and the fragment identifier points to the subordinate resource. Depending on the desired approach, different characters can be used as a fragment identifier. Practical use in the Semantic Web shows that there are two (main) different approaches to URL fragmentation. Some utilize hash-character (#) and another group of practitioners rely on slash-character (/).

From a web architecture perspective, the utilization of the hash indicates the usage of an in-document anchor. As such, all hash URIs resolve with the same namespace thus resolve to the same resource. This has the advantage that the ontology can be downloaded in one pass, but it also has the disadvantage that the file can become very big. The slash character on the other side, is used to indicate a folder-path organization in a website allowing for decentralization of the information in different folders.

Method 3 (Look up names))

For every name that has a hash namespace, make a GET call against the namespace. For every name that has a slash namespace, make a GET call against the name. The content type should be set correctly. Resolve redirects, if any. If the returned resource is an ontology, check if the ontology describes the name. If so, N is a linked data conformant name. If not, the name may be wrong.

Evaluation

The preferred way of denoting namespaces in COSI is through hash namespaces. A simple GET call on any class or object property will lead to a documentation of the ontology when the correct content-type is provided together with a description of the name, satisfying the linked data conformant name.

Conclusion: *Verified*

5.2.1.4 Method 4: Check Naming conventions

An important aspect in the paradigm of Linked Data is the well-considered URI naming strategy (W3C Working Group, 2014). A sound naming strategy will improve the understanding and reusing of the ontology in practice. In the context of the Semantic Web, interoperability is a major topic, and it is related to the diversity of formats in which knowledge resources are expressed, the differences in granularity or coverage of models, and also to the linguistic descriptions associated with semantic representations.

The used naming convention should label classes and properties that are of absolute assistance for human understanding, supporting ontology adopters in checking consistency and avoiding inaccuracies (Montiel-Ponsoda, et al., 2011). Naming convention have also been of great assistance in ontology mapping. They have also been shown to be of great assistance in tasks such as ontology mapping (Šváb-Zamazal, et al., 2008), information extraction (Müller, et al., 2004), or natural language generation (Bontcheva, 2005) (Montiel-Ponsoda, et al., 2011).

Method 4 (Check naming conventions))

A proper naming can be checked by comparing the local part of the URI with the label given to the entity or by using lexical resources like Wordnet (Fellbaum, 1998). Formalize naming conventions (like multi-word names and capitalization) and test if the convention is applied throughout all names of a namespace. Check if the URI fulfils the general guidelines for good URIs, i.e. check length, inclusion of query parameters, file extensions, depth of directory hierarchy, etc.

Note that only local names from the same namespace, not all local names in the ontology, need to consistently use the same naming convention, i.e. names reused from other ontologies may use different naming conventions.

Method 4 is related to the selection of a meaningful URIs to label classes and properties in an ontology.

Evaluation

In the development of COSI, a lexical reference is used for the evaluation of the naming conventions. The names used in COSI reflect the meaning of a class or property. As the technical documentation is based on English, the class names have been checked against

WordNet. 33.67% of the class names are directly found in Wordnet. The rest is composed of multi-words still found in Wordnet.

The naming convention is done following recommendations by (Heathe, et al., 2011). The following conventions are adopted:

1) Camel casing is used for multi-word names,

2) the names in URI are closely matched to the labels given to the entity. In case of multi-word names, the name follows convention 1 for camel casing, while the label is the normal multiword and

3) names do not contain any query parameters

Conclusion: *Verified*

5.2.1.5 Method 5: Metrics of Ontology reuse

This methodology checks the reusability factor adopted in the ontology. The method provides a metric that should show how easy it is to ease share, exchange, and aggregate information based on the ontology.

Method 5 (Metrics of ontology reuse)

We define the following measures and metrics:

- Number of namespaces used in the ontology N_{NS}
- Number of unique URIs used in the ontology N_{UN}
- Number of URI name references used in the ontology N_N (i.e. every mention of a URI counts)
- Ratio of name references to unique names $R_{NU} = N_{UN} / N_N$
- Ratio of unique URIs to namespaces $R_{UNS} = N_{UN} / N_{NS}$

Check the following constraints. The percentages show the proportion of ontologies that fulfil this constraint within the Watson EA corpus, thus showing the probability that ontologies not fulfilling the constraint are outliers.

$R_{NU} < 0.5$(79. 6%)

$R_{UNS} < 5$ (90. 3%)

$N_{NS} >= 10$(75.0%)

The baseline defined in the methodology is a result of Watson EA corpus[134]. The metric provides a reference metric for a best practice approach. Failure to comply should not necessarily indicate a bad modelling approach, rather than guide to improvements.

[134] Watson was a search engine developed by the Knowledge Media Institute (d'Aquin, et al., 2007). The complete engine indexed about 130 000 ontologies. The full indexed results of the ontology were made available for experiments and research in the Semantic Web Challenge tracks (http://challenge.semanticweb.org/).

Evaluation

COSI makes use of 20 Namespaces (N_{NS}=20) and has 137 unique URIs pointing to classes and properties defined in COSI. In total, COSI makes reference to 370 URIs.

Table 27: Metrics of Ontology Reuse

Ontology	N_{NS}	$\mathbf{N_{UN}}$	$\mathbf{N_{N}}$	$\mathbf{R_{NU} = N_{UN} / N_{N}}$	$\mathbf{R_{UNS} = N_{UN} / N_{NS}}$
COSI	20	137	370	0,37	6,85
Method 5				R_{NU}<0.5	R_{UNS} <5

Table 27 presents the metrics of namespaces, unique URIs and reference URIs found in COSI. As it can be seen, COSI complies with the recommended metrics in 2 out of 3 cases. Considering the factor or ontology reuse, COSI makes use of 20 external ontologies (N_{NS}=20), a number twice as high as the reference in Method 5. This indicates a good reuse of the ontology.

COSI also complies in the metric measuring ratio of name references to unique names, but fails to comply ratio of unique URIs to namespaces. This indicator is affected by the high number of classes and properties defined in COSI preamble. Lowering the number of properties or classes, probably a future task will generate an acceptable ratio to the recommendation of Method 5.

Conclusion: *Partially Failed*

5.2.1.6 Method 6: Check name declaration

This method checks if a declaration of a name and its type is properly declared in the ontology. The method is related to older versions of OWL, where the ontology encoding language did not require names to be declared.

This problem is addressed in OWL 2, where the language enforces declaration of

Method 6 (Check name declarations))
Check every URI to see if a declaration of the URI exists. If so, check if the declared type is consistent with the usage. This way it is possible to detect erroneously introduced punning.

Vrandecic refers to the corpus of 2008 in his research and he addresses this as Watson EA Corpus (Vrandečić, 2010).

names, so that tools can check if all used names are properly declared.

Evaluation

The feature was introduced to avoid punning in OWL. COSI relies on OWL 2 where the declarations are axioms, stating not only that a name exists but also its type, i.e. if it is declared as a class, an individual, a datatype, object or annotation property.

Conclusion: *Verified*

5.2.1.7 Method 7: Check literals and data type

Beside the object properties, ontologies rely on the use of literals that represent data values, or the so called data properties. This method checks if set of allowed data types is used.

As a standard practice, most ontologies rely on a set of data types defined by the XML Schema Definition (van der Vlist, 2002). In OWL 2, a larger range of required data types is presented to support numbers, text, boolean values, binary data, URIs, time instants etc.. Beside the standard literals contained in the XML Schema or OWL 2, developer might define their own custom data types, but based on method 7, this should be avoided whenever possible.

Evaluation

> **Method 7 (Check literals and data types))**
>
> A set of allowed data types should be created. All data types beyond those recommended by the OWL specifications should be avoided. There should be a very strong reason for creating a custom data type. `xsd:integer` and `xsd:string` should be the preferred data types (since they have to be implemented by all OWL conformant tools).
>
> Check if the ontology uses only data types from the set of allowed data types. All typed literals must be syntactically valid with regard to their data type. The evaluation tool needs to be able to check the syntactical correctness of all allowed data types.

Data types referred in COSI rely only on OWL and RDF specifications. No custom data types are defined and used.

Conclusion: *Verified*

5.2.1.8 Method 8: Check Language tag

Language tags are used as meta-properties to name classes and properties in an ontology. They state the natural language used by the literal, based on the user/client language preferences, the specific language tag will be returned referring to specific classes or properties of the ontology. This method checks on the presence of language tags with literals.

Method 8 (Check language tags))

Check that all language tags are valid with regard to their specification. Check if the shortest possible language tag is used (i.e. remove redundant information such as restating default scripts or default regions). Check if the stated language and script is actually the one used in the literal.

Check if the literals are tagged consistently within the ontology. This can be checked by counting n_l, the number of occurrences of language tag l that occurs in the ontology. Roughly, n_l for all l should be the same. Outliers should be inspected.

Evaluation

Although a multi-lingual ontology is desired, in the current version of COSI, only the English language tags are used. Language tags are used on all textual description of entities for the `rdfs:label` and `rdfs:commet` properties.

Conclusion: *Verified*

5.2.1.9 Method 9: Check labels and comments

Labels and comments in ontologies are a set of metadata that provide a human readable description on the ontology itself, the classes and object properties used.

Method 9 (Check labels and comments))

Define the set of relevant languages for an ontology. Check if all label and comment literals are language tagged. Check if all entities have a label in all languages defined as being relevant. Check if all entities that need a comment have one in all relevant languages. Check if the labels and comments follow the style guide defined for the ontology.

This method is related to the recommended practice of providing description for all the terms defined in the ontology. The descriptions should be in the appropriate language, marked with the language tag.

Evaluation

In order to improve the understanding and utilization of COSI, terms define a *rdfs:label* property that is used to provide human readable names and an *rdfs:comment* property providing a textual definition. As already expressed in Method 8, only the language tag *@en* is used and no other language is foreseen to be added in the near future.

Conclusion: *Verified*

5.2.1.10 Method 10: Check for superfluous blank nodes

Blank nodes are an RDF feature that allows representing a node or a graph without an explicit name. The practice is quite common in software developing and it is usually referred to as anonymous classes (or functions). Blank nodes can be internally referred but are not exposed to the external applications. Although the blank nodes are a standard feature in RDF, (Vrandečić, 2010) argues that *blank nodes should be avoided unless structurally necessary.*

Method 10 (Check for superfluous blank nodes))
Tables 2.1 and 2.2 (in (Vrandečić, 2010) list all cases of structurally necessary blank nodes in RDF graphs. Check every blank node to see if it belongs to one of these cases. Apart from these, no further blank nodes should appear in the RDF graph. All blank nodes which are not structurally necessary should be listed as potential errors.

Evaluation

Blank nodes are not part of the ontology definition per sè, but they might be created during the population and creation of graphs based on the ontology definition. For this reason, this method does not apply to the modelling of COSI. Nevertheless, the use of blank nodes is still a valid practice.

Conclusion: *Inapplicable*

5.3 Syntactic Quality

Evaluation of quality with regard to syntactic aspects is associated to the formal style and the way the ontology is written. This aspect evaluates the syntax that is used to serialize the ontologies. There are several serializations option related to an ontology. As it can be imagined, the syntax aspect is related to issues such as comment style, XML validation, and the creation of XML Schema etc. Vrandečić lists only one method under this aspect.

5.3.1.1 Method 11: Validating against an XML Schema

An ontology can be implemented in a specific description logic and then expressed or serialized using different serialization formats. Although the OWL 2 has its own language encoding, the serialization and publication of the ontology in RDF/XML is an expected outcome.

Although RDF/XML is an XML based serialization, there are a few syntactical approaches that are recommended and evaluated by this method. First, the serialization should generate a valid XML document. The XML validation should be performed on an ontology to verify its conformance to the serialized syntax ontology on which it is built. The serialized file should make use of RDF-style comments rather than XML-style comments. And second, qualified ontologies adopted by the Semantic Web community should be used.

Method 11 (Validating against an XML schema))

An ontology can be validated using a standard XML validator under specific circumstances. In order to apply this, the ontology needs to be serialized using a pre-defined XML schema. The semantic difference between the serialized ontology and the original ontology will help in discovering incompleteness of the data (by finding individuals that were in the original ontology but not in the serialized one). The peculiar advantage of this approach is that it can be used with well-known tools and expertise.

Evaluation

Evaluations regarding the validity of the XML Schema were performed over the RDF/XML Serialization of the ontology. Validation check is tested through the RDF Validator (Prud'hommeaux, 2006). Although a RDF/XML serialization of COSI exists (and is published on the documentation page), this is not the primary encoding file expressing the ontology.

Conclusion: *Verified*

5.4 Structural Quality

Structural aspects of an ontology are widely explored in comparison with the previous aspects. This is also due to the fact that the structure of an ontology is a graph representation, and graphs are well researched in the field of mathematics. With regard to ontologies, several measures are proposed to analyse the inner structure of an ontology. (Vrandečić, 2010) points that there are more than forty different metrics that may be used to measure the structure of an ontology. Due to simple implementation, most ontology toolkits provide ready access to a number of these metrics. Sometime, based on these metrics, ontology repositories provide annotations and altering options of the ontologies with regard to their structural quality.

This section is dedicated to a set of methods used for the evaluation of the structure of an ontology.

5.4.1.1 Method 12: Ontology Complexity

This method is focused on the structural complexity of an ontology. Complexity is a standard evaluation measure of an ontology language and OWL 2 is no difference. Nonetheless, the modellers of an ontology, can influence the complexity of an ontology by the features that they decide to use in the ontology.

Based on the set of features that are defined, there are different complexities of a specific ontology, such as COSI in our evaluation. The ontology language merely defines an upper bound of a possible complexity.

To define the complexity of the ontology, this method considers that the sum of all the expressivity features included in the modelled ontology should still be decidable. Tools such as *Complexity of reasoning in Description Logics* (Zolin, 2013) provide with a practical information if the complexity of the modelled ontology is within expected boundaries.

Method 12 (Ontology complexity))

We define measures counting the appearance of each ontology language feature. We do this by first defining a filter function O_T: $O \rightarrow O$ with T being an axiom or an expression type. O_T returns all the axioms of axiom type T or all axioms having an expression of type T.

We can further define a counting metric N_T: $O \rightarrow N$ *as* $N_T\ (O) = |O_T\ (O)|$.

We also define $N(O) = |O|$.

We can then further define a few shortcuts, derived from the respective letters defining DL languages, for example:

- **Number of subsumptions** $N_{SubClass}O\ f\ (O) = |O_{SubClassOf}(O)|$: the number of subsumption axioms in the ontology
- **Number of transitives** $N_{TransitiveProperty}(O)$: the number of properties being described as transitive
- **Number of nominals** $N_O(O)=N_{OneOf}(O)$: the number of axioms using a nominal expression
- **Number of unions** $N_{UnionOf}(O)$: the number of axioms using a union class expression

- etc.

With these numbers we can use a look-up tool such as the description logics complexity navigator (note: (Zolin, 2013)). If $N_O > 0$, then the nominals feature has to be selected, if $N_{TransitiveProperty} > 0$ we need to select role transitivity, etc. The navigator will then give us the complexity of the used language fragment (as far as known).

We further define $H(O){:}O \rightarrow O$ as the function that returns only *simple subsumptions* in *O*, i.e. only those *SubClassOf* axioms that connect two simple class names.

Evaluation

A list of functions expressing the if a certain type of logica axiom is found in COSI:

- $N_{SubClassOf}$ $(O) > 0$
- $N_{EquivalentClasses}$ $(O) > 0$
- $N_{DisjointClasses}$ $(O) > 0$

- $N_{SubObjectPropertyOf}$ $(O) > 0$
- $N_{EquivalentObjectProperties}$ $(O) = 0$
- $N_{InverseObjectProperties}(O) > 0$
- $N_{DisjointObjectProperties}(O) = 0$
- $N_{FunctionalObjectProperty}(O) > 0$
- $N_{InverseFunctionalObjectProperty}(O) > 0$
- $N_{TransitiveObjectProperty}(O) = 0$
- $N_{SymmetricObjectProperty}(O) = 0$
- $N_{ReflexiveObjectProperty}(O) = 0$
- $N_{IrrefexiveObjectProperty}(O) = 0$
- $N_{ObjectPropertyDomain}(O) > 0$
- $N_{ObjectPropertyRange}(O) > 0$
- $N_{SubPropertyChainOf}(O) > 0$

- $N_{SubDataPropertyOf}(O) > 0$
- $N_{EquivalentDataProperties}(O) = 0$
- $N_{DisjointDataProperties}(O) = 0$
- $N_{FunctionalDataProperty}(O) > 0$
- $N_{DataPropertyDomain}(O) > 0$
- $N_{DataPropertyRange}(O) > 0$

- $N_{ClassAssertion}(O) = 0$
- $N_{ObjectPropertyAssertion}(O) > 0$
- $N_{DataPropertyAssertion}(O) = 0$
- $N_{NegativeObjectPropertyAssertion}(O) = 0$
- $N_{NegativeDataPropertyAssertion}(O) = 0$
- $N_{SameIndividual}(O) = 0$
- $N_{DifferentIndividuals}(O) = 0$

- $N_{AnnotationAssertion}(O) > 0$
- $N_{AnnotationPropertyDomain}(O) > 0$
- $N_{AnnotationPropertyRangeOf}(O) > 0$

With these specifications, the complexity of COSI is ALCRIQ(D) within the capabilities of OWL 2. Concept satisfiability based on the tool *Complexity of reasoning in Description Logics* (Zolin, 2013) is *NExpTime,Decidable.*

Conclusion: *Verified*

5.4.1.2 Method 13: Searching for Anti-Patterns

This method is related to a set of anti-pattern. Although the naming might be confusing, an anti-pattern is what a modeller might believe to be patterns, but in fact turns to be an invalid pattern and solution. This method points to two specific well-known anti-patterns.

Method 13 (Searching for Anti-Patterns))

SPARQL queries over the ontology graph can be used to discover potentially problematic patterns. For example, results to the following queries have been found to be almost always problematic.

Detecting the anti-pattern of subsuming nothing:

```
      select ?a where {
        ?a rdfs:subClassOf owl:Nothing .
      }
```

Detecting the anti-pattern of skewed partitions:

```
      select distinct ?A ?B1 ?B2 ?C1 where {
            ?B1 rdfs:subClassOf ?A .
            ?B2 rdfs:subClassOf ?A .
            ?C1 rdfs:subClassOf ?B1 .
            ?C1 owl:disjointWith ?B2 .
       }
```

Evaluation

The defined patterns in method 13 are checked through a SPARQL query to COSI to verify the inclusion or exclusion of certain patterns. None of the above anti-patterns were found COSI.

Although this guarantees that the ontology has no problem with the defined anti-patterns, the metric is still fragile as there might be other unexplored anti-patterns.

The lack validity for the anti-patterns defined in method 13 indicates that COSI inherits no such problems.

Conclusion: *Verified*

5.4.1.3 Method 14: OntoClean meta-property check

OntoClean (Guarino, et al., 2002) is a methodology for ontology evaluation based on the formal analysis of classes and their subsumption hierarchy. It makes use of four meta-properties: rigidity, unity, dependency and identity that are applied to ontology classes to measure the adequacy of the otology by analysing the taxonomic relationships present in the ontology.

(Vrandečić, 2010) points that although the evaluation with OntoClean is expensive, it is recommended to make use of automated tools that allow an analysis based on the OntoClean methodology. One of the tools is AEON[135] (Völker, et al., 2005) an approach to automatize OntoClean checks. Some additional tools include ODEClean for WebODE (Fernández-López, et al., 2002) and last OntoEdit (Sure, et al., 2002). Unfortunately, none of these tools is found to be currently maintained (AEON lacks proper documentation, while the rest of the tools can no longer be found).

Method 14 (OntoClean meta-property check))

An ontology can be tagged with the OntoClean meta-properties and then automatically checked for constraint violations. Since the tagging of classes is expensive, we provide an automatic tagging system AEON (http://ontoware.org/projects/aeon/).
All constraint violations, i.e. inconsistencies in the meta-ontology, come from two possible sources:

- an incorrect meta-property tagging, or
- an incorrect subsumption.

The evaluator has to carefully consider each inconsistency, discover which type of error is discovered, and then either correct the tagging or redesign the subsumption hierarchy.

Evaluation

As it was not possible to use any of the tools referred in the description of this methodology, a manual evaluation was considered. The formal notions of OntoClean; rigidity, essence, unity and identity were used to tag certain meta-properties in a subset of classes from the ontology[136]. Checks were done to see whether subsumptions constrains were held. While applying the method, no problems were found.

Conclusion: *Verified*

5.5 Semantic Quality

So far, methods related to *vocabulary, syntactic* and *structural aspects* of an ontology are discussed. These aspects are all related to the semantic aspect, that are concerned with the relation between identifiers and concepts modelled in the ontology. This section is focused on a set of methods related to the semantic quality of an ontology. Methods described in this section deal with

[135] https://code.google.com/archive/p/aeon-project/

[136] Seven main classes, documented in the ontology preentation in Section 4.2 were used in the OntoClean checks.

- *normalization* or metrics used to reduce data redundancy;
- *stability* of the ontology considering real word scenarios, future needs for evolvement and axiom alternations;
- *language completeness* or the ratio between the knowledge that *can* be expressed and the knowledge that is stated in the ontology.

5.5.1.1 Method 15: Ensuring a stable class hierarchy

Metrics related to *stability* may be measured by considering the open world assumption into account. As ontologies are expected to be implemented and used in the Web, stability metrics check on conditions that the ontology need to fulfil in any situation. An example for such a metric is the difference between the longest subsumption path of the ontology, against a *stable minimal depth of the ontology,* which is a reference to the smallest number of levels the ontology class hierarchy will have no matter what axioms and individuals are added.

This method checks the ontology hierarchies (incorporating the semantic aspect) to determine whether they are stable or not.

Method 15 (Ensuring a stable class hierarchy))

Calculate a normalized class depth measure, i.e. calculate the length of the longest subsumption path on the normalized version of the ontology *md(N(O))*. Now calculate the stable minimal depth of the ontology $md^{min}(O)$.

If *md(N(O))* <> $md^{min}(O)$ then the ontology hierarchy is not stable and may collapse.

Evaluation

This metric is related to the length of the subsumption hierarchy, or else the number of levels the class hierarchy has against the stable minimal depth of the ontology. Stable metrics are metrics that take the open world assumption into account. The stable minimal depth of the ontology $md^{min}(O)$ is calculated on the instances of an ontology. Considering that the instances populating COSI are deriving from a well-defined automation process, the ontology hierarchy was found to be stable in the evaluation. To be noted: Ensuring the stability of a class hierarchy through references to an open world assumption is a challenging metric! An evaluation following the normalization of the ontology and comparing the normalized class depth versus the stable minimal depth of the ontology was also done in a small scale of the ontology (population of the ontology with 100 individuals on the main classes) and the result satisfies the requirements of method 15.

Conclusion: *Verified*

5.5.1.2 Method 16: Measuring language completeness

This method measures the language completeness of the ontology. *Language completeness* or the ratio between the knowledge that *can* be expressed and the knowledge that *is* stated in the ontology.

Method 16 (Measuring language completeness)

We define a function ϒ with the index *i* being a language fragment (if none is given, the assertional fragment is assumed) from an ontology O to the set of all possible axioms over the signature of O given the language fragment *i*. We introduce C_i as language completeness over the language fragment *i*.

$C_i(O) = (|\{X|X \Upsilon (O),O |= X \vee O |= \neg X|) / |g(O)|$

Evaluation

Evaluating the metric is not feasible. The method has no baseline for a comparison, and although a language completeness value C_i can be calculated based on specific population of the ontology, the number will differ based on different population scenarios.

Conclusion: *Inapplicable*

5.6 Representation quality

Quality evaluation related to the representation aspect deals with semantics of the ontology and how these semantics are structurally represented on the terminology and naming conventions used in an ontology. Methods listed under this aspect aim to identify mistakes that may arise between the formal specification and the conceptualization of the ontology. This section is dedicated to a set of methods used for the evaluation of the vocabulary of an ontology.

5.6.1.1 Method 17: Explicitness of the subsumption hierarchy

This method ascertains the explicitness of the subsumption hierarchy. Similar to Method 15, this method evaluation is based on the maximum depth of the taxonomy, referred as *T*.

It also makes use of the maximum subsumption path length referred to as *SL*. The metric introduced in this method is the *explicitness of the subsumption hierarchy* denoted by *ET(O)* and calculated as *ET(O) =TD(O)/SL(O)*, where O is the ontology (COSI).

Method 17 (Explicitness of the subsumption hierarchy))

Calculate *ET(O).*

- If *ET(O)* = 1 everything seems fine
- If *ET(O)* < 1 then some of the classes in the ontology have collapsed. Find the collapsed classes and repair the explicit class hierarchy
- If *ET(O)* > 1 part of the class hierarchy has not been explicated. Find that part and repair the class hierarchy

Evaluation

With a subset of 100 individuals on the COSI, the *ET(O)* is computed and the following measures are obtained

$$ET(COSI) = 5/5 = 1.$$

Per definition of the metric, If *ET(O)* = 1 everything seems fine, and there is a balance in the taxonomy hierarchy and the semantics.

Conclusion: *Verified*

5.6.1.2 Method 18: Explicit terminology ratio

This method is based on assessing that the ratio between classes and class names defined in the ontology (and also property and property names) is always equal to 1.

This method is inspired by the measure called the *Class / relations ratio* (Gangemi, et al., 2005). The original method would return the ratio between classes and the relations in the ontology graph. From a representation aspect of an ontology, Vrandečić points that there should be an evaluation metric based on the ratio between each of the two components, i.e. the ratio of classes and class names $R_C(O) =|C_N(O)|/|C(O)|$ and the ratio of properties and property names $R_P(O) =|P_N(O)|/|P(O)|$.

As stated on the method description, if $R_C(O) = R_P(O) = 1$ then the representation of classes and properties in the ontology is sufficient and correct.

Method 18 (Explicit terminology ratio))

Calculate RC(O) and RP(O).

- If $R_C(O) = R_P(O) = 1$, this indicates no problems with the coverage of elements with names in the ontology
- If $R_C(O) < 1$ *or* $R_P(O) < 1$ and the ontology does not include a mapping to an external vocabulary, this indicates possible problems since a number of names have collapsed to describe the same class
- If $R_C(O) < 1$ *or* $R_P(O) < 1$ and the ontology includes a mapping to an external vocabulary, we can remove all axioms providing the mapping and calculate $R_C(O')$ and $R_P(O')$ anew
- If $R_C(O) > 1$ *or* $R_P(O) > 1$, this indicates that not all interesting classes or properties have been given a name, i.e. the coverage of classes and properties with names may not be sufficient

Evaluation

With regard to COSI:

$R_C(O) = |C_N(O)|/|C(O)| = 65/ 65=1$

and

$R_P(O) = |P_N(O)|/|P(O)| = 73/73=1$

The ratio between the normalized and not normalized ontology graph remains the same.

Conclusion: *Verified*

5.7 Contextual Quality

Quality evaluation related to the contextual aspect is associated with artefacts and conditions that influence the present state of an ontology. The aspect can be seen as a direct relation of the content and usefulness of the ontology for the users as well. A classical metric for the evaluation of this aspect are competency questions. Such questions describe what kind of knowledge the resulting ontology is supposed to answer. This section is dedicated to a set of methods used for the evaluation of the contextual quality in an ontology. It will start with competency questions as mentioned, and gradually moves on to methods related to more technical criteria that analyse the contextual aspect.

5.7.1.1 Method 19: Checking competency questions against results

In Section 4.1, discussing on the methodology followed to achieve the classes and properties that are present in COSI, we discussed the need for competency questions as questions at a conceptual level that the ontology needs to answer. A list of competency questions was also listed in

Table 7: Competency **Question**. This method verifies the adequacy of an ontology using competency questions.

Evaluation

Based on competency questions defined in Section 4.1.4 *Steps involved in the development of COSI*, a set of SPARQL queries have been executed on a populated version of COSI and the result satisfied the expected outcome

Method 19 (Checking competency questions against results))

Formalize the competency questions as a SPARQL query. Write down the expected answer as a SPARQL query result, either in XML or in JSON. Compare the actual and the expected results. Note that the order of results is often undefined.

Conclusion: *Verified*

5.7.1.2 Method 20: Checking competency questions with constraints

This method is a follow up on the previous method. It is considered useful not in the verification of the state or quality of the existing ontology, but rather an insight on the capabilities to extend and amend the ontology with newer axioms and conditions. While the previous method dealt with the ability of the ontology to answer successfully the competency questions, in this method, the competency questions are used to generate new ontologies (by using `SPARQL CONSTRUCT`) resulting in a method referred to as competency questions with constraints.

Method 20 (Checking competency questions with constraints)

Formalize the competency questions for ontology O as a SPARQL CONSTRUCT query that formulates the result in RDF as an ontology R. Merge R with O and a possibly empty ontology containing further constraints C. Check the merged ontology for inconsistencies.

Evaluation

This method is not considered in the evaluation of the actual formalization in COSI. As described, this method might be helpful for ontologies that are highly dynamic and face constant changes. Iterations and versions of COSI made use of previous methods to assess the internal quality.

Conclusion: *Inapplicable*

5.7.1.3 Method 21: Unit testing with test ontologies

Unit tests are a software testing methodology that relies on the presence of computer programs that execute individual units of code in order to determine whether these code snippets are fit for use or not. This method ports the same experience in the evaluation of an ontology.

Method 21 (Unit testing with test ontologies))

For each axiom A^{+}_{i} in the positive test ontology T^{+} test if the axiom is being inferred by the tested ontology O. For every axiom that is not being inferred, issue an error message. For each axiom A^{-}_{i} in the negative test ontology T^{-} test if the axiom is being inferred by the tested ontology O. For every axiom that is being inferred, issue an error message.

Evaluation

In the older iterative versions of COSI, Protégé plug-in OWL Unit Test framework[137] was used. This plugin is outdated and can no longer be used with the newer versions. In any case, as pointed out by Vrandečić, test ontologies are meant to be created and grown during the maintenance of the ontology.

Conclusion: *Verified*

5.7.1.4 Method 22: Increasing expressivity

Ontologies in information systems often need to fulfil the requirement of allowing reasoners to quickly answer queries with regards to the ontology. The performance of an ontology reasoning engine is related to the expressivity of the ontology. This method tents to introduce a metric of checks in case of expressivity increase. The ontology is checked how it behaves in the presence of a peculiar case of expressivity increase.

Method 22 (Increasing expressivity))

An ontology O can be accompanied by a highly axiomatized version of the ontology, C. The merged ontology of $O \cup C$ has to be consistent, otherwise the inconsistencies point to errors in O.

Evaluation

COSI makes use of the HermiT reasoning engine (Shearer, et al., 2008). Hermit is used as an evaluation engine as well. It can determine whether the ontology is consistent, identify subsumption relationships between classes, assess on expressivity and more. Hermit

[137] The original documentation page for this plugin is no longer found. The code can still be found at http://smi-protege.stanford.edu/svn/owl-unit-test/. Accessed February 2016

is used to verify all the axiomatic triples implemented in COSI and check for potential problems related to increasing expressivity. During verification, no inconsistencies were reported.

Conclusion: *Verified*

5.7.1.5 Method 23: Inconsistency checks with rules

Another metric of evaluating an ontology would be to introduce logical constrains in the ontology and check if they are respected in a practical implementation. The scenario included in Method 23 relates to porting the ontology to a logic programming language like *Datalog* (Grosof, et al., 2003) and checking if the constrains introduced are present.

> **Method 23 (Inconsistency checks with rules))**
> Translate the ontology to be evaluated and possible constraint ontologies to a logic program. This translation does not have to be complete. Formalize further constraints as rules or integrity constraints. Concatenate the translated ontologies and the further constraints or integrity constraints. Run the resulting program. If it raises any integrity constraints, then the evaluated ontology contains errors.

This method checks the presence of inconsistencies in an ontology with the help of rules.

Evaluation

As mentioned earlier, the Hermit reasoner is used to validate COSI and identify inconsistencies, if any. The Hermit reasoner covers quite well verification of inconsistency checks with rules (and the authors of this engine are also involved in the development of the old Datalog tool). During verifications with Hermit, no inconsistencies were reported.

Conclusion: *Verified*

5.8 Evaluation in the context of an application and a task

Evaluation of an ontology in the context of an application and a task is related to the adequacy of the ontology to accomplish its purpose. While the evaluation procedures discussed so far are based on a set of methodologies that asses the correctness of the ontology per sé, the evaluation of an ontology in an application and a task addresses concerns on practical usage of the ontology. In Section 4.3 – discussing on Sheer Curation, it was presented the eSciDoc Browser as a tool were COSI is used natively. Beside eSciDoc

Browser, COSI could easily be used within different tools such as Wisski[138], or Vivo. VIVO is a product addressing organization of publications and other documentation of academic personnel, but core component of VIVO is a VITRO (Duraspace, 2015), a generic ontology editor. Vitro can easily be adapted to load other ontologies and it was tested with COSI ontology as well. Another interesting product is Wisski[139], a Drupal based set of modules that may be used as a Virtual Research Environment (VRE) and makes use of ontological data models. COSI can easily be loaded in Wisski as well and make use of the extended Drupal modules to allow population of data based on COSI.

Arguing that the use of COSI is also possible in the case of practice task related to the conservation of information gathered during a scientific investigation, below are a list of projects where earlier iterations of COSI were used.

BW-eLabs - Information Networking in the Scientific Research Process

BW-eLabs[140], a project involving 7 partners from the federal state of Baden Württemberg. The project was initiated with the goal of advancing heterogeneous experimental resources (remote and virtual) for sustainable coverage. Special focus of the project was the use of raw data and experiments for research and advanced education.

The use case partners in BW-eLabs are the *Freiburg Materials Research Center* (FMF)[141] and *Institute of Applied Optics at Stuttgart University* (ITO)[142]. The final raw output of investigations in these institutes differs. In case of FMF, the objects are mainly absorption and photoluminescence spectra. Experiments conducted by ITO provide mostly digital holograms. Despite the different output results, the BW-eLab workflow was generalized to be represented in the BW-eLabs Ontology[143], an extension of the CSMD, the predecessor of COSI. In a general perspective, both laboratories produce calibration and configuration information, which are important for correctly understanding and interpreting the captured data from instruments. Data objects created in the laboratory are annotated and registered in a

[138] http://wiss-ki.eu/

[139] http://wiss-ki.eu/ Wisski is a German acronym for "Wissenschaftliche KommunikationsInfrastruktur", which can be translated as "Scientific Communication Infrastructure"

[140] http://www.bw-elabs.org/index.en.html

[141] FMF - Freiburg Materials Research Center - https://www.fmf.uni-freiburg.de/index_en.html/

[142] Institut für technische Optik - http://www.uni-stuttgart.de/ito/index.en.html

[143] BW-eLabs Ontology has been cited in Ontological formalization of scientific experiments based on core scientific metadata model (Brahaj, et al., 2012) as an extension of CSMD, an older verison of COSI.

repository at the very moment they come into existence, providing a first working functionality of our sheer-curation implementation.

eKinematix - Virtual research environment with integrated information structure

eKinematix aimed to provide a platform for the organization of investigative and creative activities in the field of mechatronics and robotics. The aim of the project was to provide a virtual research environment where researchers could execute their investigations and benefit from automated documentation on the process. The virtual research environment would document the creation of new artefacts from the conception moment, up to the publication of technical papers describing the fully contextual information and provenance of the artefacts created.

The project anticipated a platform that would facilitate the enhancements of the documentation process by linking information across different knowledgebases, a concept shared by COSI as well. The aim of the project was the improvement of the reutilization of research and development results with focus on R&D in mechatronics. The raw output was usually stored as XML data[144] and was generated by a specific setup of the eKinematix design software. The project was executed successfully and future implementations of it are in progress.

NanoCollect - Collecting and Organising Knowledge on Nano-materials

NanoCollect is a vision project aiming to organize knowledge information on nano materials. The use of nano materials is of increasing relevance for the industry. Beside the beneficial exploitation of nano materials, there are critical concerns on the safety of nano materials and the respective technology. Therefore, analyses of potentially hazardous properties as well as exposure scenarios gain more and more importance with regard to protection of users and environment. Within the last decade, the amount of data from investigations dealing with the properties of nano materials and their biological effects has immensely increased. Research on nano particles is handled across many institutions spread around the globe. With regard to the difficulties of spotting hazardous properties in these particles, management and coherence of potentially hazardous properties are crucial for the scientific community, standardization authorities like OECD and ISO/CEN as well as producers of nanoproducts. As science nowadays is increasingly collaborative and different movements such as open science are aiming to lower the barriers of sharing scientific knowledge between researchers and institutions, the focus of NanoCollect is to focus on a

[144] The XML in this project is related to the raw-data, results of the research activity, and not to the utilization of XML Schema for modelling of XML Data. The modelling is still relying on RDF & COSI.

solution that will provide broad accessibility to nanomaterial data and contribute to releasing valuable and trusted scientific information as linked open data. NanoCollect envisions a healthy ecosystem of decentralized data providers, based on ontology engineering with focus on security and nano particles. The involvement of FIZ Karlsruhe in the project proposal was based on the experience of the institution with semantic data modelling and the potential integration of COSI in the workflow of investigations for the domain and the alignment of results from different repositories and institutions. A project proposal for NanoCollect was not selected for funding under an EU grant, but the project remains an interesting vision on how information on nano safety can be organized in a linked data world.

As a conclusion, a set of projects supporting documentation and annotation of information outcome from scientific investigation were presented. The project's range covered different disciplines but COSI could be applied with the same simplicity to each of the cases, being a strong indicator for the compatibility of COSI for different types of investigations.

5.9 Summary of Ontology Evaluation

In the chapter, the ontology evaluation framework shaped by (Vrandečić, 2010) was used to evaluate COSI. Vrandečić framework on ontology evaluation is an incremental work based on different evaluation methodologies. The framework assesses an ontology based on eight criteria: *Accuracy, Adaptability, Clarity, Completeness, Computational-efficiency, Conciseness, Consistency* and *Organizational fitness*. As direct assessment of these criteria is not easy, 23 evaluation methods are defined and applied for the evaluation of the ontology at hand. These methods are grouped in six aspects of an ontology. In this chapter, each of the methods was tested on COSI and a quantification metric ranging from *Verified, Inapplicable, Deferred, Failed* was used to indicate successful compliance to the expected outcome. The summarized values of these predefined metrics are shown in Figure 29: Summary of COSI Evaluation.

Nineteen of the twenty-three methods are found to comply as ***Verified*** and they meet the expectations of the evaluation methodology used. Four methods were not found to be verified. Out of these methods, only three methods are found ***Inapplicable*** for the current state of COSI. As an example, the check for superfluous blank nodes (Method 10) is not applicable for the ontological formalisation, but may be relevant in case of a population of the ontology. In a similar case, method 16 *Measuring language completeness* is not feasible for evaluation as the method has no baseline for a comparison and method 20 *checking competency questions with constraints* is helpful only for ontologies that are highly dynamic and face constant changes. While there are no ***Deferred*** methods, there is one method that is marked as partially failed. Method 5, *Metrics of Ontology reuse* recommends a set of reference values that are found to be shared among some prominent ontologies. Out of three references in this method, only one of the reference values is outside the recommended range. Although it is possible to adapt the ontology reuse or the number of internal classes and properties in COSI, the reference values of method 5 indicate a best practice. Failure to comply with the reference does not indicate failure in ontology functionality, but rather a partial failure to comply to a best practice scenario, as observed by (Vrandečić, 2010). Section 5.8 presented a set of projects where COSI was applied.

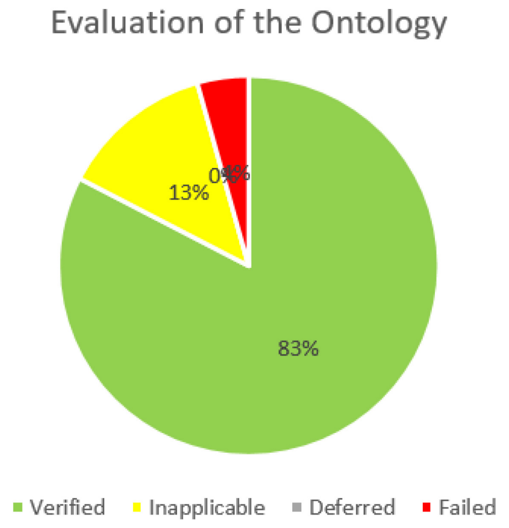


Figure 29: Summary of COSI Evaluation

Conclusively, based on the above summarized evaluation of the COSI, it can be stated that the ontology represents the required quality standards, and possesses the expected structural and semantic characteristics.

Chapter Summary

The objective of this chapter was to evaluate COSI to ensure that the developed ontology is of an acceptable quality. The evaluation is done by following the framework presented by (Vrandečić, 2010). This framework is based on the evaluation of 23 methods that are based on six aspects of an ontology. Results found by the evaluation of each of the methods are presented in a chart to provide a summary of the COSI evaluation. In addition, Section 5.8 presents evaluation of COSI in the context of an application and a task as required by domain stakeholders.

"Alles ist Wechselwirkung"
Alexander von Humboldt (von Humboldt, 2003)

6. Discussion

The sophistication of information technology introduced by advancements in computing technology has reshaped the way scientific investigations and in a broader concept, the way research activity is executed. Computers and access to digital equipment have shaped the way we interact and perceive solution to problems and opportunities presented. Science and research activities make no exceptions from the mind shift change we are facing. Visionary publications such as the Jim Gray's "*The Fourth Paradigm*" (Gray, 2009), already predicted how science will be transformed in the new information technology era. Vast amount of scientific effort will go to the analysis and interpretation of data generated from research, transforming and creating new prospects in research domains. Other flagship strategy publications such as "*Riding the Wave, How Europe can gain from the rising tide of scientific data*" a document defining the digital agenda for the European Union (Wood, et al., 2010) stresses the importance of research infrastructures that support collaboration beyond the traditional borders. This thesis is inspired by such predictions and is based on requirements for simple and yet powerful methods that can influence the research activity.

This chapter is dedicated to a discussion on the findings, additional contributions, applications and implications of this work. The findings of the research are summarised and discussed in Section **6.1**, *Research Findings*. Section **6.2** continues with a brief description of the *Significance of the Results* and Section **6.3** points to some *Limitations of the Approach*. Portability of the model to other science disciplines is discussed in Section **6.4**, *Portability to Other Scientific Disciplines*. The last section is dedicated to this research implication and the final conclusion.

6.1 Research Findings

Focus of this research is the exploration of a novel approach on supporting research data curation by developing a method and defining an automated data curation process where data can be easily annotated. The preservation of research results goes in parallel with the execution of investigations. Preserved results should contain a broad spectrum of metadata needed to understand the data, and also incorporate long time preservation information that allows these data to be cited and re-used by other interested parties. Central to the thesis is the development of a model to formalise scientific investigations targeting mainly the structural sciences, a subset of Life Sciences and Natural Sciences.

The investigation of these research questions led to the following main contributions:

Contribution I: *Formalisation of an ontological model for the representation research accomplished in the course of a scientific investigation.*

This contribution is attained through a theoretical model and artefacts, which are summarize below:

- The presentation of conceptual model for the definition of scientific investigation activities.

  The conceptual model is used to establish the Core Ontology for Scientific Investigations which represents a conceptualization of knowledge from scientific investigations.
- The model is well documented and published online. Documentation of the ontology is also provided in Appendix C
- I demonstrate that the proposed formalization is practicable and a real-world scenario is presented to prove its applicability.

Contribution II: *Implementation of a generic solution, referred to as sheer curation, an activity that allows automation of metadata creation and ingests on investigation execution. The data are marshalled through COSI, a semantic model. In addition, through semantic enrichment, the data gathered can be related and referenced to other research results.*

This contribution is based on the specific formalization of the model presented in Contribution I and a well-defined workflow. The contribution is achieved following:

- A data management curation process is presented introducing a novel solution to address semi-automatic data enrichment. The process is presented coupled with the ingest activities in a Virtual Research Environment.
- Through evaluation and experiments with real-world datasets from various sources, I demonstrate that the above solution is valuable and feasible.

Findings Summary

The formalization of an investigation through COSI allows referencing of all core entities and offers a vivid presentation of the investigation arrangement. The amount of data describing a specific investigation can be used to explain the final result set. With the increasing influence of information technology in our daily work, these data need to be reusable and well curated. And yet, researchers need to spend as little time as possible in annotating and providing the necessary metadata for the final result-sets. This research deals exactly with these issues and in addressing them, the following results were achieved.

1) A set of core entities is evidenced that are can be used to best describe a scientific investigation
2) In order to address the vision of interoperability, a group of modelling techniques were evaluated and OWL2 (and RDF serialization) was assessed as the most suitable choice. The same modelling technique allows additional operations of reasoning and satisfies criteria defined in the requirements of this thesis
3) An information model was formalised to express information related to scientific investigations and research results in the *Core Ontology for Scientific Investigations*
4) Addressing issues of proper metadata annotation, a practical procedure to incorporate the generation of metadata on the execution of investigations is defined. This process of a semi-automated curation activity (sheer-curation) removes the burden of populating results of investigations with abundant metadata

In the first step toward the engineering of COSI, the required core entities needed for the representation of scientific investigation metadata are evidenced. These core entities are required to document the base information needed for the understanding of data generated in the course of a scientific investigations. The examination for the core entities of a scientific investigation has focus on research activity in disciplines of structural science such as nano technology, chemistry and mechatronics (see Use Cases and Applications for applied uses in Section **5.8**). A central notion for the analysis was the entity of *investigation*. The other entities are discussed in relation to *investigation*, and follow a trail that lead to *results*, another key entity in COSI. The analysis undertaken in evidencing the main entities used to describe an investigation is based in documentation provided by partner stakeholders in technical reports mainly from project BW-eLabs, but also through colleagues and community feedback. The entities gathered and used in COSI present investigation information from the initial *Research Question*, to *Hypothesis*, concrete *Investigation* and up to resulting *Publication*. The automated curation workflow allows for the publication of all this information in a simple way.

In order to address the research questions and challenge the thesis main claim, a suitable information modelling technology is needed. Data models have a strong influence on the way data are processed. Through the course of scientific research, novel data models have impacted the progress of information technology. By encapsulating an abstraction layer and hiding technical details, models provide a translation of simplified scenarios from real world problems. In the scope of this work three modelling techniques are evaluated. The considered models are *Hierarchical Models* where XML is the most prominent representative, *ER Models* that are usually represented by RDB systems and the younger *RDF Model*, a graph structure that can be used to encode among others, Description Logic definitions. These 3 models were evaluated based on the following criteria:

1) *Logical to physical dependence*, an evaluation of how the logical representation of these structures depends on the physical implementations. In this case, the Hierarchical Models faced some restrictions in consideration to the other 2 models evaluated.
2) *Data Structure organization* was evaluated with respect to the internal data organization. As expected the Hierarchical Model (tree hierarchy) and ER Model (flat tables) are less advanced than the graph model used in RDF.
3) *Linking capabilities* were more restrictive in the case of Hierarchical Models, where links were possible only within the scope of the current document. Improvements existed in the ER Model with links that refer within the database structure and almost no limitations in the case of RDF Models, where URI linking is a first class citizen.
4) *Scalability* placed the ER Model as the most prominent choice with the other two models following.
5) *Practical use* evaluation was assessed with consideration to the practical use in industry and academia. In fact all three models are well recognized and advocated for use in these domains
6) *Intended use* evaluation focused on the intended use of the data models. For the considerations within this research work, it was found that the RDF Model is more appropriate.

7) *Support for cross reference checking* was also evaluated with special attention to the requirements of interoperability, where RDF Model provides better support through federated queries.

From the analysis of the criteria, the RDF Model asserts to be a more suitable candidate for modelling metadata and empowering interoperability and reasoning capabilities on the annotated date. The criteria used were discussed in Section **3.2.5**, *Assessment of Modelling Techniques* and a visual representation is presented in Table **6** on the same section.

Based on the two first outcomes of the research, the process of evidencing core entities and an analysis on an appropriate modelling technique, the next result of this research is the formalization of an ontology named Core Ontology for Scientific Investigation. The ontology was modelled in OWL2 and Description Logic. Beside the core entities, it contains a set of properties, attributes that present the way the entities relate to each other. The ontological presentation is the main contribute of the thesis, presenting a novel way on how the investigation data can be presented. The definition of the metadata in OWL2 and the presentation in RDF serialization opens new possibilities in data analysis and data centric activities.

Data modelled in OWL allow for different analytical and interoperability capabilities and fulfil the requirements gathered in this dissertation[145]. Annotation of the result-data with semantic technologies, at their point of creation can produce abundant contextual information to allow reproduction of the investigation and provide a clear understanding of the process, increasing trust in the research processes.

The proposed number of entities contained in COSI is limited, just as the real entities in an investigation environment are limited, yet requiring researchers to annotate all the result data and relate them to the investigation environment is resource consuming. To show that the developed ontology is viable in practical scenarios, a workflow on how the model can be incorporated in a virtual research solution is presented. The result is the definition of a curation activity that happens on the flow, as the investigations are executed.

[145] Requirements of this thesis are evidenced in Section 3.2 and 3.2.6

6.2 Significance of the results

I'm facing rather too many silly and unnecessary barriers to my research:
lack of data sharing, lack of online data availability
lack of data in an immediately machine-readable digital format[146]

While we are facing a data deluge (Hey, et al., January 2003), there are still barriers that prohibit researchers from easily finding and sharing results from other peers (Dallmeier-Tiessen, 2011). The frustration of missing on the benefits of collaboration and alignment of result researches is related partially to policies related to commercialization concerns and partially to lack of appropriate infrastructures that facilitate data sharing. This work is not concerned on policy making or widely advocated initiatives such as *open access*, but rather on improvements related to data modelling and exploration of techniques that improve generation of valuable research data with as little resources as possible. In the presented results, two main aspects are covered.

Data Model

First is the provision of a novel approach to data modelling that allows integration of plentiful metadata that can be treated as *logic* concepts and not merely as literals. These concepts are defined in an ontology, which allows among other actions, inference and reasoning operations. Such capabilities simplify the relation of metadata with other metadata in other repositories that rely on the same formalization technique. Adhering to this specific formalization, we not only comply with requirements of publishing data in machine-readable format, but also provide information in machine-understandable format.

Data Annotation

The second aspect of this dissertation is related to facilitation of metadata annotation on the fly. This process occurs while investigations are executed. The annotated research data contains a full spectrum of metadata describing the investigation they derive from, and other social and technical context. This contribution affects topics of data-sharing and online data availability. The findings show that through use of sheer-curation, it is possible to publish the research results and the necessary set of metadata needed for proper comprehension of such data. By *publication*, I refer to a proper publication activity, where data are stored in a repository system and equipped with the necessary preservation features such as *persistent identifiers*, *versioning history*, *licencing information* and *data access control mechanisms*.

[146] The snippet was extracted by an online article of Ross Mounce, a biology-graduate advocating for Open Science. The original citation was retrieved on September 2012 from http://rossmounce.co.uk/aboutme/

Initiatives such as Nanopublication[147] already place focus on research publications that are not fully academic publishing, but minimal triples that provide a valid research assertion. Such vision fits to the requirements of many scientists, to have access to raw data they can consume, reuse and cite in their research. There is clearly a shift from the traditional citation referring classical scholar communications such as journal or conference papers, to references of raw data that reside on (publicly) accessibly repositories. This thesis outlines a concrete approach on how research data can be published online and be treated as citable research results.

Data Science and Data Reuse

Published research data does not need to be only positive data, or data that have fulfilled the prediction of the initial hypothesis. Negative data, or data that did not prove the intended investigation hypothesis still contain valuable information for other researchers, especially considering the growth of computational power available to even home users (and progress of citizen science[148]).

Data science, an activity that finds presence in many different discipline, is based on research being executed based on analysis and interpretation of data. With the evolution of digital equipment and data science, raw data generated in research facilities across the globe can be reused. The term stands for the modern abundance of digital data from many sources that can be mined with clever software for discoveries and insights. Its promise is smarter, data-driven decision-making in every field. The field known as "big data" offers a hot topic in research and practical implementations, although the term is sometime used in a strict literal understanding. It is not the amount of data in single repository that defines "big data", but rather the vast amount of repositories, may these repositories be small or large ones, and their contribution to manageable data (on the net). From a research perspective, big data bottlenecks lay in the lack of software that automates the gathering, cleaning and organizing disparate data, which are plentiful but disorganized. Organizing and aligning these data fits in an actual handcrafted work. Data scientists, *according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labour of collecting and preparing unruly digital data* (Lohr, 2014), before these data can be explored for useful research. As enormous value in data science comes from combining different data

[147] Not to be confused by the use of Nano in the name, Nanopublication is related to acknowledgement of scientific pucblicationpublication of is the smallest unit of publishable information: an assertion about anything that can be uniquely identified and attributed to its author.

[148] Citizen Science is scientific research conducted, in whole or in part, by amateur or nonprofessional scientists. Citizen science is having a momentum with the increased data online as part of open access or open data initiatives. In other cases, citizen science is executed through valuable observations of common citizens. To the core of its activities, are the data and flexible data solutions that allow them to exchange information.

sets, it is necessary to consider modelling techniques that allow more than simple data serialization (such as XML). Data from sensors, documents, the web and conventional databases all come in different formats. In the course of this dissertation, it was argued that in order to match such demands, a modelling effort such as the one resulting in COSI can be used through relation to ontologies based on Description Logic and OWL2 to support the annotation of metadata. This formalization can be expressed in different encodings, including RDF/XML, a format that provides a simple backward compatibility with XML-familiar repositories. How the choice of the application of this new technology affects data repositories is discussed briefly in *Section 6.5* discussing on Repositories of the Future.

The application of semantic technologies; the modelling of COSI as a core ontology in scientific investigations and the presentation of the semi-automated curation process coined as sheer-curation ascertain the hypothesis of this thesis. It is possible to improve the publication process and the quality of research data by formalizing the metadata as an ontology. The publication process of research results is improved by merging COSI with a semi-automated curation process. The formalization of the information in an ontology allows for data alignment and reasoning operations that improve and facilitate data mining in data-science. Annotation of the result-data with semantic technologies, at their point of creation can produce abundant contextual information that allow reproduction of investigations and provide a clear view of the process increasing trust in the research processes. In addition, the use of semantic technologies to annotate the research-data increases their visibility and usability.

Data centric activities are already an important factor in research and industry. Those who have access and the proper digital setup to mine data are already in a better position to dwell into innovative quests. (But) It seems that the prospects information management can open to new knowledge acquisition are still underestimated. Data and data preservation are seen in many cases as mere storage of information that backs a research and whose existence is related solely to this goal. The concept of data quality in many technical standards is related to the data longevity perception. Raw data are almost considered as something dead; artefacts whose memorial needs to be preserved for the sake of an historical argument. Data reuse is a prominent topic in many scholar communications, but few stress the importance of challenging this topic from a technical perspective. The hypothesis of this thesis questions the availability of a modelling technology that treats data as *living* information. Such data are still valuable; they can be queried and associated with other data from other repositories to provide information beyond the original creation goal. It is through a novel modelling technique that analytical and statistics operations can exploit new value and make re-use a necessity. Solutions that envision and facilitate data annotation, knowledge extraction and data re-use do impact and benefit researchers with little information on information technologies.

Ontological formalisations are not a novel approach anymore. The State of the Art section already referred to some research similar to the topic of this research by grouping them in scientific investigation data modelling initiatives, such as Core of Scientific Metadata Modelling (Matthews, et al., 2010) and data modelling ontological formalisation efforts such as ontologies of OBI Foundry, EXPO (Soldatova, et al., 2006) or CIDOC-CRM (International Organization for Standardization, 2014)**.** COSI stands in the intersection of these initiatives, bringing the ontological formalization, already applied in different scenarios of data management to the formalization of scientific investigation. It improves the formalization of prior modelling efforts based on hierarchical structure and presents the power of Description Logic to data management efforts related to scientific investigation.

6.3 Limitations of the approach

Semantic Technologies are no longer considered a new frontier. The topic is almost 20-year-old nowadays, and a lot of extensive research has been dedicated to the topic during these years. Innovators such as Tim Berners-Lee, prominent companies such as Google, Yahoo and Bing have pushed toward use of semantic technologies, pointing to data management benefits once these technologies are applied. And still, the progress of semantic technologies and their penetration in the daily operations is slow.

In evaluating limitations of the presented research, the first drawback is related to the limited presence of semantic implementations in our day to day operations. While I discuss that, there are substantial benefits of applying the formalisation of COSI in practice, the statement is true only with consideration that ontological formalisations and large corpora of data encoded in ontological representation exist. In other words, the benefits of the presented solution are best exploited once this same technology becomes mainstream and is widely adopted in research data management practices. This limitation is mainly related to the added value of semantic technologies and the ability to interlink data and concepts across different graphs or repositories. Although this is a disadvantage of the modelling technology, the disadvantage exists only within the vision of semantic technologies and not to the actual state of data modelling technologies. In other words, the utilization of an ontology model has no disadvantage over an XML Schema or other modelling technologies already discussed. Representation of investigation data in semantic technology has no disadvantage over existing models; it distinguishes itself in the vision of an interconnected word of information.

Due to the decentralized and linked architecture of the envisioned semantic web (or Web of Data), answering queries requires accessing and combining information from multiple repositories or graphs. The application of semantic technologies in the Web of Data has resulted in large corpora of interlinked datasets from diverse realms. Recent statistics on linked data confirm the presence of 85 billion triples available from more than 3500 different datasets (LODStats, 2015). Although this is a large number of data expressed in RDF, the community considers it a very small number. Following the idea of Linked Data, there is an enormous potential for integrated querying over multiple distributed data sources. Any application that relates information from more than one data source needs to execute queries over various sources at the same time. These queries are referred as *federated queries*. From a user perception this means that data from multiple distributed sources can be queried transparently as if residing in the same database. In order to join information provided in multiple sources, expensive computing resources are required. The expensiveness of the federated queries is related to the extent of the graphs merged in the query process. During tests done in the course of this work, the performance of federated queries deteriorates drastically based on the size of the graphs merged[149]. While the expensiveness of federated queries seems a rather serious problem, there is extensive research such as (Schwarte, et al., 2011) (Görlitz, et al., 2011) and different techniques such as caching or RDF Stream Processing. It is to be foreseen that the increasing attention of the industry toward semantic technologies will encourage solutions to the problem of federated query performance as well.

6.4 COSI Portability to Other Scientific Disciplines

COSI has been developed as formalism with a focus on scientific investigations performed in structural sciences such as Nano technology, chemistry, material science, earth science, biochemistry etc. Although it was never the aim of this research to provide a meta-model to address all scientific investigation practices, it is interesting to see how the model relates to research in other science disciplines and if the model can be ported to address similar needs in other disciplines. In this section I assess the alignment of COSI with some prominent formalisation used across different research disciplines. The evaluation does not intend to advocate COSI as replacement ontology for these specific ontologies or the disciplines they are bound to, but rather evaluate if the model can be ported and used in other scenarios. Also, it is interesting to evaluate the compatibility degree of COSI with these formalisms with consideration of the desired abstraction level. For the comparison, the following models are considered:

a) *The Common European Research Information Format (CERIF)*,
b) *VIVO ontology*, part of the VIVO project that aims to enable a network of scientists and

[149] All tests in this thesis are done through the open source triple stores Sesame (http://rdf4j.org/) and Kiwi (http://marmotta.apache.org/kiwi/triplestore.html)

c) *CIDOC Conceptual Reference Model (CIDOC-CRM)*, a model that provides definitions and a formal structure for describing concepts and relationships used in cultural heritage documentation.

CERIF (Asserson, et al., 2002) is a data-centric data model that provides information on research entities, their activities and their output. The model provides high flexibility with formal relationships. The scope of the model is broad and interdisciplinary. The initiative was born as part of the EU Working Group on Research Databases and was transferred to the euroCRIS CERIF Task Group (Cordis EU, 2015). euroCRIS has led CERIF through various upgrades and extensions since 2000 and the latest version of CERIF is 1.5. Beside the XML-Schema representation, there exists also formalism in OWL (Jörg, 2013). The actual version of CERIF has a richness of 293 entities and 1814 attributes. The model is meant to provide an extensive coverage of research activity compared to the abstract level desired in COSI. Beside technical information, CERIF supports a wide range of entities including information on human resources such as qualifications of the investigators, personal curricula or administrative information such as funding of the project etc.

VIVO (Krafft, et al., 2010) is a very interesting project aiming to document researcher's activity within one or more institutions. VIVO has a practical implementation in a usable product that supports curation of the data through the VIVO ontology. The project aimed to enable a "National Networking of Scientists" aiming to connect the researchers in US (Krafft, et al., 2010), but recently VIVO has gained a lot of visibility in the international community as well. The VIVO ontology is developed to provide points of access for a virtual researcher's community. It aims to organize and present information about people, research, and their research activities. VIVO instances are encouraged to be installed in research institution to provide documentation on the research and publications of each of the staff of the organization. The actual version of VIVO has integrated concepts from other domain ontologies such as BFO and this has increased the number of entities in the ontology to more than 400, although the initial native entities comprise a smaller number. VIVO is considered in this evaluation due to its application in different disciplines to represent research activity.

The last model to be included in this evaluation is CIDOC-CRM (International Organization for Standardization, 2014). CIDOC, or the *Comité International Pour La Documentation*, through the International Council for Museums has contributed to a number of standards in documenting concepts and relationships used in cultural heritage documentation. The CIDOC CRM aims to promote a common understanding of cultural heritage information by providing a documentation framework that any cultural heritage information can be mapped to. The initiative's ambition is to provide a common language for domain experts serve as a guide for good practice of conceptual modelling. The model is centred on 94 entities and 168 attributes. Although the number of entities differs significantly, this metric will not be considered in the discussion. In most cases, this number represents in depth extension of concepts in each information model. With consideration of COSI as a CORE ontology, it is important to have a good coverage on the top-level concepts and possible on their direct predecessors. Deeper levels represent in many cases specific domain related entities. For example, in COSI we use the Activity entity to connect an Entity, an Agent and an Activity in a general way; in CIDOC-CRM the same entity would be aligned with the entity Period and its extensions that are used to represent specific use cases of curation needs. It is to be noticed that the CIDOC-CRM provides a wide range of entities that are related not only to the description of cultural items, but sometime cover preservation references such as Identifier Assignment, Symbolic Object and also many periodical entities.

To discuss the alignment capabilities of each of the ontologies, with regard to scientific investigation, a set of prominent entities that carry most of the information needed in representing an investigation (and as one might imagine, these are the main properties in COSI or the uppermost[150] entities in the hierarchical tree) are assembled. Possible porting entities from the other 3 ontologies are evaluated to see possible alignments. Results are presented in Table 28.

Table 28: Entity alignment in COSI, VIVO, CERIF and CIDOC-CRM.

COSI	VIVO	CERIF	CIDOC-CRM
Investigator (Roles)	Person	Person (cfPers)	E21 Person E39 Actor
Instrument	Equipment	Equipment (cfEquip)	E18 Physical Thing E24 Physical Man-Made Thing
Institution	Organization	Organisation	E74 Group
Facility	Address	Facility	E45 Address E53 Place
Motivation Publication	Article	Publication	E31 Document
Result Publication	Article	Result Publication	E31 Document / E34 Inscription

[150] Since most concepts in an OWL ontology derive from the Thing entity, the *uppermost* term used here will denote to the level 2 of the hierarchy.

Result-Set	Document/Dataset	Result Product (cfResProd)	E1 CRM Entity
Investigation	n/a	n/a	E29 Design or Procedure E87 Curation Activity
Parameters	n/a	n/a	E57 Material
Study	Project	Project	E87 Curation Activity
Hypothesis	n/a	n/a	n/a

All the considered ontologies contain references to the central figure of an active *investigator*, an individual involved in a research process. In the case COSI, the entity is further sub-classed with attention to the involvement of the agent in the concrete investigation, eg: *Principal Investigator, Observer* etc. In VIVO, the same entity is sub-classed to indicate a person's academic appointment to a specific faculty of a university or institution of higher learning, eg: *EmeritusProfessor*, *PostDoc* etc. The *Person* entity is not further classified under CERIF and CIDOC-CRM. The *Actor* entity in CERIF allows relation of a person to skills, CV and other attributes allowing extended information. The same entity in CIDOC-CRM is constrained to relations to cultural items with no further context on the entity itself. Basic information on the *Actor* are inherited by the *Person* class.

An *Instrument* represented as devices used in the course of investigations in COSI can be aligned with the *Equipment* entities in VIVO and CERIF. The *vivo:Equipment* is in fact an entity deriving from a BFO hierarchy. CERIF as well, inherits the definition of this entity from WordNet, although it extends the definition by a number of attributes that relate more to the use of *cosi:Instrument*. The closest entities relating to an instrument in the CIDOC-CRM are the *E18 Physical Thing* and *E24 Physical Man-Made Thing*, although by definition an instrument does not necessarily restrain in a "Man-Made thing". *E18 Physical Thing* on the other side has a generalized scope.

Definitions of *Institution* and *Facility* have a well aligned representation across these four ontologies, with a slight derivation in the naming of CIDOC-CRM, where an *Organization* is denoted as a *Group*, although the definitions align seamlessly. *Motivation* and *Result Publication* entities in COSI can be aligned to an *Article* in VIVO and to a *Publication*, *Result Publication* in CERIF. In case of CIDOC-CRM, the alignment is with yet another broad concept, a *Document*. In some cases, this might be aligned with an *Inscription*, defined as short texts attached to any instances with the aim of documenting artefacts.

The *Result-Set* defined in COSI is best aligned with the CERIF *Result Product*. In case of VIVO and CIDOC-CRM, we should rely on broader entities such as *Document* and a *CRM*

Entity. For the key entity of *Investigation* in COSI, it was not able to find a proper alignment in the other evaluated ontologies. The VIVO ontology had a few references extended BFO concepts relating to *Case Studies* and *Interventional Study*, but these are a subset of the broader concept of *Investigation*. With regards to the specific domain of application, in CIDOC-CRM this entity can be aligned with a *Curation Activity* or a *Design or Procedure.* In a similar fashion, *Parameters* in COSI cannot be aligned with a proper match in VIVO and CERIF, but can be related to the *Material* entity in CIDOC-CRM due to the nature of the domain of application for CIDOC-CRM. The last entity evaluated, *Study* finds a similar match in VIVO and CERIF entity *Project* since the use of the concept *study* is used as an aggregator of investigations organized under a specific plan.

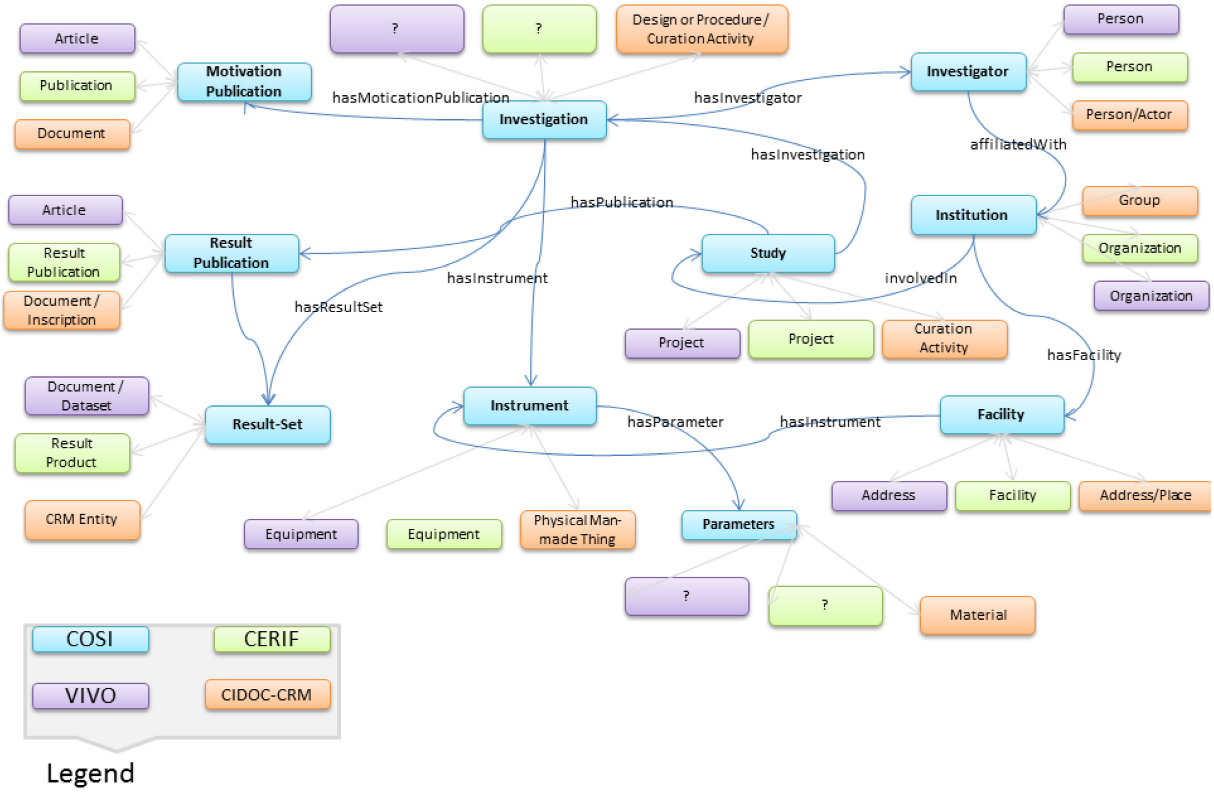


Figure 30: Alignment of key COSI entities and some prominent relations. Please see Figure 18 for an accurate relation of COSI entities.

So as we see, the other information models considered have a good coverage in representing the research domain. CERIF and VIVO seem to align better, while most concepts of CIDOC-CRM are too broad. This is also related to the scope of these information models. While CERIF and VIVO aim to represent the research activity, CIDOC-CRM is more related to curation activity for museums and other forms of archiving. Two of the main properties of COSI, the *Investigation* and the *Parameters* influencing an investigation are missing in the alignment table (these are part of the subset analysed, other entities such as *Principal Investigator, Calibration* etc. are not discussed as too detailed). From the analysis, we can deduct that COSI covers a missing space in the information models used for the specific domain of information management. In attempt to support interoperability and integration of the information models discussed, COSI can be used to domains of application of CERIF and VIVO, but will eventually not be fit for the too specific CIDOC-CRM application. From the representation in Table 28, we can see that the properly mapping the information semantics represented in CERIF, VIVO and COSI is a modest challenge, especially considering the OWL support in entity alignment.

6.5 Implications

The motivation for this thesis was driven by recent developments on information technology and unexploited opportunities in data management or curation activity related to research activities. The thesis promotes a novel research data curation process and an information modelling methodology that will both influence the way we interact with research data and disseminate research findings. The selected modelling methodology is based on a novel information management technique, part of a new vision for the web and information exchange that will eventually disrupt the way traditional research findings are published and shared.

Findings advocated on this thesis are all related to technical proposals, but as it usually happens, improvements in technology lead to implications on the social activities and once both these domains are affected positively, new standards are defined. From the technological perspective, the application of COSI together with the presented sheer-curation activity will drastically influence the generation of machine process able result-sets and at the same time simplify research data management by including more contextual information for less effort spent by researchers. Similar positive consequences are to be expected in empowering data mining and other analytical activities that open new opportunities for research and data sciences. These technological implications will gradually influence the social perceptions on research data management. The simplification of result data publishing and the exploited benefits of data mining will (gradually) impose new standards of scholarly communication exchange as well. Three key implications of the application of the advocated solution are discussed below.

Scholarly Communication

In *Foundations*, while discussing the backgrounds of this research, we stopped briefly on the history of scholarly communication. The current practice was traced back to

Oldenburg's *Philosophical Transactions of the Royal Society,* a journal born in the 17th century. Oldenburg's allowed for an authoritative medium that allowed researchers to publish their research findings in reports adhering to the scientific method. The detailed description of research activity was crucial for the acceptance and evaluation in the *Philosophical Transactions*. To allow reviewers and other researchers to reach confidence in the research reports, abundant information on the result sets and other contextual information was included. With advancements in information technology in 20th and 21st century, the results cited in similar research reports have significantly grown. The traditional research reports published in academic journals, conferences and other activities under the umbrella of scholar communication need to relate to concrete data sets that can be found online, assessed with regard to trust in published reports and in best cases reused for further research. These result sets need to be accessible online, citable and findable the same way the research reports are.

In the solution presented in this thesis, a specific information modelling methodology based on semantic technologies is presented together with a curation activity that annotates the results with metadata from their moment of the creation is discussed. Generated datasets may be directly published in a repository and yet benefit from the preservation features of academic reports such as persistent identifier, versioning history and references to additional metadata. As these data may be reusable in other investigations, or need to be cited in different papers, we will eventually see a shift toward (massive) publication of result sets; researchers will be accredited for publication of their research data beside the traditional research reports. Authority, a key element of trust in research community may gradually lose terrain in favour of direct validity checks on the validity of acclaimed results. This will of course happen in those scenarios where full information on the reproducibility of investigations that led to the result data can be provided. In any case, application of COSI in practical implementations, or other ontologies that may borrow or have similarities with COSI will create a new practice in the scholarly communication, a practice based on result datasets published, cited and accredited for. The importance of result-sets being published will gradually be valuated as important as report papers. Author-level metrics that attempts to measure productivity and citation impact of the publications of scholars will be tweaked to reflect reusability of research result-sets beside other scholar communications.

Research Data Interoperability

Considering the information management progress in the last decades, there is immense amount of information related to research activities in repositories accessible on the internet. Attention toward open access, open data and other open frontier policies are encouraging a barrier-free exchange of research information. On the other side, research data are located at different repositories, in different data formats and different access interfaces.

Although there is massive information related to research in different disciplines on the net, most of the data is scattered and unstructured. Data sets in each of these repositories are not interlinked with each other, a challenged addressed by an analytical process referred to as data mining.

Data mining is defined as the practice of examining large existing knowledge base repositories in order to generate new information. Technology related to data mining deal with extraction of concepts or meaningful hints from the data examined and matching operations of these data to already known concepts. The complexity level of data mining operations is related to the inner structure and the machine readability of the examined data. The resources needed and the qualities of the mining operations correlate with the inner structuring of data, and the ability of machines to process these data. In case of semantic technologies, data are easily interlinked among each other through ontology alignment that makes effective discovery, mechanization and assimilation possible. This is mainly due the fine machine readable format utilized by semantic web. Such data can be easily shared and processed by automated services as well as people.

COSI presents a novel solution where information on result-sets derived from research investigations are modelled in OWL. This information and can be expressed in RDF, stored in RDF serialization formats and queried through SPARQL. Storing such information in machine readable and process able formats, lowers the resources needed for data mining operations and increases the quality of the generated information. Employing COSI or similar or deriving information models will greatly influence the interoperability of information across disciplines. In such cases, it would be easy to relate findings from one discipline with the application in another one. In an illustration example, examiners involved in cosmetics, sunscreens or special clothing production will be able to cross-search for information on a special element, example: zinc oxide nanoparticles to find from a Nano-Safety knowledge base that the particle also produces substantial DNA damage.

The application of COSI in real world examples has the potential to improve the interoperability of information across different disciplines, a much-desired result nowadays (see *Riding the Wave – How Europe can gain from the rising tide of scientific data* (High Level Expert Group on Scientific Data, October 2010)).

Repositories of the Future and Digital Libraries

Digital libraries have evolved dynamically over the past two decades, and so has the use of the term. While there are different definitions for digital libraries, their definitions relate to four main concepts: *Collection, Service, Organization* and *System* (Brahaj, et al., 2013). As I encourage for new modelling technique and technology in this thesis; it is interesting to see how this model will affect digital libraries or repositories of the future.

Table 29: Genera of digital libraries grouped in four groups, based on **(Brahaj, et al., 2013)**.

Collections	**Services***	**Organization**	**Systems***
- Collections - Organized collections - Managed collection - Focused collection - Electronic resources - Collection of collections - Organized collection of digital resources - Collection of information objects - Collection of documents in electronic format - Resource	- Library services - Dynamic federated structures - Information storage - Retrieval systems - Distributed environment - Collection of services - Group of services	- Organization - Operational organization -Socio-technical systems - Virtual organization	- Systems - Tools - Electronic resources - Database on hypertext environment - Environment - Library - Socio-technical systems - Networks of technology

From the four main concepts deducted in (Brahaj, et al., 2013), the embrace of the proposed modelling technique will greatly affect definitions and evaluation of digital libraries with respect to those defined as *Services* and *System* concepts. Support for ontological modelling and operations will impact the need for more computational capabilities for digital libraries, besides the existing more relaxed storage, uptime or input-output capabilities. Some large projects have already invested in repository solutions that support ontologies. Europeana[151] an internet portal that acts as an interface to millions of cultural objects that have been digitised throughout Europe already contains information modelled in EDM (Doerr, et al., 2010), an ontological model. The backend of Europeana supports perfectly SPARQL queries, allowing cross references not only within Europeana but other repositories such as DBPedia[152]. While Europeana is a pioneer in the embracement of this new technology, application of COSI in practice or other models based on it will lead to repositories providing and having crucial part of their ecosystem the support for semantic technologies. Support for this layer will influence the capabilities and the evaluation of such digital libraries from a technical and features perspective. Digital libraries will no longer be considered as storage organizations, but as live information processing machines that allow a

[151] Europeana.eu is an EU initiative - www.europeana.eu

[152] DBpedia is a crowd-sourced community effort to extract structured information in RDF from Wikipedia and make this information available on the Web – www.dbpedia.org

dynamic interface to knowledge extraction across different sources, fulfilling in part the vision of Vannevar Bush in Memex (Bush, 1945). Bush envisioned Memex as a machine that would allow access to a giant knowledge base system that would get information through electromechanical controls, books, microfilm cameras and readers, all integrated into a *large desk. Most of the microfilm library would have been contained within the desk, but the user could add or remove microfilm reels at will* (Bush, 1945). This vision resembles with the federated query in SPARQL where search is possible to be done across repositories and graph inclusions. Embracement of COSI and similar semantic information modelling, and support for such models in practice will change the way we perceive repositories in the future.

7. Bibliography

Abdallah Samer A. und Ferris. Bob The ordered list ontology 0.72. [Online]. - 2010. - 03 2015. - http://purl.org/ontology/olo/core.

Apache Foundation Apache Commons [Online]. - 2016. - 02 2016. - https://commons.apache.org/.

Asserson Anne, Jeffery Keith G. und Lopatenko Andrei CERIF: past, present and future: an overview. [Journal]. - 2002.

Asunción Gómez-Pérez, Fernández-López Mariano und Corcho Oscar Ontological Engineering [Buch]. - [s.l.] : Springer, 2004.

Baader Franz und Werner Nutt Basic description logics [Journal] // In Description logic handbook . - 2003. - S. 43-95.

Bachman Charles W. Data structure diagrams. [Journal] // ACM Sigmis Database 1. - 1969. - Bd. 2. - S. 4-10.

Ball Andre How to License Research Data [Bericht]. - Edinburgh : DCC How-to Guides, 2014.

Basic Formal Ontology (BFO) bfo - Basic Formal Ontology (BFO) - Google Project Hosting [Online]. - 2015. - 02 2015. - https://code.google.com/p/bfo/.

Berners-Lee Tim Linked Data [Online]. - 7 2006. - 07 2014. - http://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee Tim Relational Databases on the Semantic Web [Online]. - W3, September 1998. - 07 2014. - http://www.w3.org/DesignIssues/RDB-RDF.html.

Berners-Lee Tim, Hendler James und Lassila Ora The semantic web [Journal] // Scientific american. - 2001. - 5 : Bd. 284. - S. 28-37.

Bézivin Jean und Gerbé Olivier Towards a precise definition of the OMG/MDA framework [Konferenz] // Proceedings. 16th Annual International Conference on Automated Software Engineering (ASE 2001). - 2001. - S. 273-280.

BFO - Basic Formal Ontology BFO - Basic Formal Ontology [Online]. - University of Saarland, 2015. - 02 2015. - http://ifomis.uni-saarland.de/bfo/.

Bizer Chris, Cyganiak Richard und Heath Tom How to publish linked data on the web [Journal]. - 2007.

Blei David M. Hierarchical Models [Bericht]. - 2011.

Boehringer David [et al.] LiLa–Library of Labs [Artikel]. - 2010.

Bontcheva Kalina Generating tailored textual summaries from ontologies. [Journal] // In The Semantic Web: Research and Applications. - 2005. - S. 531-545.

Brahaj Armand Capturing and Sharing Scientific Research Data [Journal] // Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies. - 2012. - ACM. - S. 31.

Brahaj Armand Core Ontology for Scientific Investigations [Online]. - 04. 10 2015. - 04. 10 2015. - http://dx.doi.org/10.6084/m9.figshare.1564549.

Brahaj Armand Vocabulary of Scientific Topics [Online]. - 2016. - https://dx.doi.org/10.6084/m9.figshare.3406594.

Brahaj Armand, Matthias Razum und Schwichtenberg Frank Ontological Formalization of Scientific Experiments Based on Core Scientific Metadata Model [Journal] // In Theory and Practice of Digital Libraries. - 2012. - Springer Berlin Heidelberg. - S. 273-279.

Brahaj Armand, Razum Matthia und Hoxha Julia Defining Digital Library [Journal] // In Research and Advanced Technology for Digital Libraries. - 2013. - Springe Berlin Heidelberg. - S. 23-28.

Brahaj Armand, Razum Matthias und Schwichtenberg Frank Ontological formalization of scientific experiments based on core scientific metadata model [Konferenz] // Theory and Practice of Digital Libraries. - 2012.

Brank Janez, Grobelnik Marko und Mladenic Dunja A survey of ontology evaluation techniques [Journal] // In Proceedings of the conference on data mining and data warehouses (SiKDD 2005). - 2005. - S. 166-170.

Broekstra Jeen, Kampman Arjohn und Van Harmelen Frank Sesame: A generic architecture for storing and querying rdf and rdf schema [Konferenz] // In The Semantic Web—ISWC ,. - 2002.

Buneman Peter und Davidson Susan B. Data provenance – the foundation of data quality [Journal]. - 2013.

Bunge und Mario Treatise on Basic Philosophy (Volume 4): Ontology II, A World of System [Buch]. - Boston : Reidel Pub. Co., 1979.

Bunge und Mario Treatise on basic philosophy(Volume 1): Ontology I: the furniture of the world. Vol. 1 [Buch]. - [s.l.] : Springer, 1977.

Bush Vannevar As We May Think [Artikel] // The Atlantic Monthly. - July 1945.

Casey Marshall Jarsync: a Java implementation of the rsync algorithm [Konferenz]. - 2013.

Chen Hai und Tompa Frank Wm. Set-at-a-time access to XML through DOM [Artikel] // In Proceedings of the 2003 ACM symposium on Document engineering (DocEng '03). ACM. - 2003. - S. 225-233.

Chen Peter P., Thalheim Bernhard und Wong Leah Y. Future directions of conceptual modeling [Buchabschnitt] // Conceptual Modeling. - [s.l.] : Springer Berlin, 1999.

Chen Peter Pin-Shan The entity-relationship model—toward a unified view of data [Journal] // ACM Transactions on Database Systems (TODS) - Special issue: papers from the international conference on very large data bases. - 1975. - Bd. Volume 1 Issue 1. - S. 9-36.

Cordis EU CERIF: the Common European Research Information Format [Online]. - Cordis EU, 2015. - 3. August 2015. - http://cordis.europa.eu/cerif/.

Cornell University Library/Research Department Moving Theory into Practice: Digital Imaging for Libraries and Archives [Online]. - Cornell University Library/Research Department, 2003. - 25. 04 2015. - https://www.library.cornell.edu/preservation/tutorial/metadata/table5-1.html.

Crawford Susan und Stucki Loretta Peer review and the changing research record. [Journal] // Journal of the american Society for Information Science. - 1990. - 3 : Bd. 41. - S. 223-228..

Cycorp OpenCyc | Cycorp [Online]. - Cycorp, 2015. - 02 2015. - http://www.opencyc.org.

Dallmeier-Tiessen Suenje “Drivers and Barriers in sharing research materials – in HEP and beyond” [Bericht]. - [s.l.] : HU-Berlin, 2011.

d'Aquin Mathieu [et al.] Watson: Supporting next generation semantic web applications [Konferenz]. - Vila real, Spain : WWW/Internet conference, 2007.

De Nicola Antonio, Missikoff Michele und Navigli Roberto A software engineering approach to ontology building [Journal] // Information systems. - 2009. - 2 : Bd. 34. - S. 258-275.

Delbru Renaud [et al.] A Node Indexing Scheme for Web Entity Retrieval [Konferenz] // Proceedings of the 7th Extended Semantic Web Conference (ESWC). - 2010.

DFG DFG Classification of Subject Area, Review Board, Research Area and Scientific Discipline (Status: 06/2008) [Online]. - Deutsche Forschungsgemeinschaft, 06 2008. - 07 2014. - http://www.dfg.de/download/pdf/dfg_im_profil/gremien/fachkollegien/dfg_fachsystematik_en_08_11.pdf.

Dictionary.com Dictionary.com's 21st Century Lexicon [Online]. - 2015. - 03. 02 2015. - http://dictionary.reference.com/browse/axiomatization.

Doerr Martin [et al.] The europeana data model (EDM) [Journal] // World Library and Information Congress: 76th IFLA general conference and assembly. - 2010. - S. 10-15.

Douglas Harper Online Etymology Dictionary [Online]. - Douglas Harper, 2001. - 25. 04 2015. - http://www.etymonline.com/index.php?allowed_in_frame=0&search=meta&searchmode=none.

Duraspace Vitro Architecture [Online]. - Duraspace, 2015. - 20. October 2015. - https://wiki.duraspace.org/display/VIVO/Vitro.

e-IRG Data Management Task Force Report on data management [Journal]. - 2009.

EUR-Lex Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [Bericht]. - 1996 .

European Commission Commision Recommendation of 17.7.2012 on access to and preservation of scientific information [Journal]. - 17.7.2012.

Farrar Scott und Langendoen D. Terence An OWL-DL Implementation of Gold. [Buchabschnitt] // Linguistic Modeling of Information and Markup Languages. - [s.l.] : Springer Netherlands, 2010.

Fellbaum Christine WordNet: An Electronic Lexical Database (Language, Speech, and Communication) [Buch]. - [s.l.] : MIT Press, 1998.

Fernández-López Mariano und Gómez-Pérez Asunción The integration of OntoClean in WebODE. [Journal] // CEUR Workshop Proceedings. - 2002.

Franz Incorporated SPARQL API reference [Online]. - Franz, Incorporated, 2014. - 07 2014. - http://franz.com/agraph/support/documentation/v4/sparql-reference.html.

Gabillon Alban und Letouzey Léo A View Based Access Control Model for SPARQL [Journal] // 4th International Conference on Network and System Security (NSS). - 2010. - IEEE. - S. 105-112.

Gangemi Aldo [et al.] Ontology evaluation and validation: an integrated formal model for the quality diagnostic task [Bericht]. - 2005.

Gangemi Aldo [et al.] Sweetening ontologies with DOLCE [Artikel] // Knowledge engineering and knowledge management: Ontologies and the semantic Web. - 2002 : Springer Berlin Heidelberg, 2002. - S. 166-181.

Gauch Hugh G. Scientific Method in Practice [Buch]. - Cambridge : Cambridge University Press. p. 45, 2003.

Giereth Mark On partial encryption of rdf-graphs. [Journal] // The Semantic Web–ISWC . - 2005. - Springer Berlin Heidelberg. - S. 308-322.

Gil Yolanda [et al.] Provenance XG Final Report [Bericht]. - http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/ : W3C, 2010.

Giunchiglia Fausto und Zaihraye Ilya Lightweight ontologies [Buchabschnitt] // Encyclopedia of Database Systems. - [s.l.] : Springer US, 2009.

Godik Simon [et al.] OASIS eXtensible access control 2 markup language (XACML) [Journal] // Technical Report OASIS. - 2002.

Goerz Günther, Oischinger Martin und Schiemann Bernhard An implementation of the CIDOC conceptual reference model (4.2. 4) in OWL-DL. [Konferenz] // In Proceedings of CIDOC. - 2008.

Gómez-Pérez Asunción Ontology evaluation [Buchabschnitt] // In Handbook on ontologies. - [s.l.] : Springer Berlin Heidelberg, 2004.

Görlitz Olaf und Staab Steffen Federated Data Management and Query Optimization for Linked Open Data [Journal] // New Directions in Web Data Management. - 2011. - Springer.

Gray Jim [et al.] Scientific Data Management in the Coming Decade [Bericht]. - [s.l.] : Microsoft, 2005.

Gray Jim The Fourth Paradigm: Data-Intensive Scientific Discovery [Buch] / Hrsg. Hey Tony, Tansley Stewart und Tolle Kristin.. - Redmond, Washington : Microsoft Research, 2009. - S. 284. - 978-0-9825442-0-4.

Grosof Benjamin [et al.] Description Logic Programs: Combining Logic Programs with Description Logic [Journal] // Proceedings of the Twelfth International World Wide Web Conference, WWW2003. - 2003. - S. 48-57.

Grotendorst Theresa Semantische Wissensrepräsentation im Forschungsdatenmanagement - Metadaten als Linked Data - Erstellung einer Ontologie für das Projekt [Buch]. - Karlsruhe : Hochschule Karlsruhe, 2011.

Gruber Thomas R A Translation Approach to Portable Ontology Specifications [Journal] // Knowledge Acquisition. - 1993. - 2 : Bd. 5. - S. 199-220.

Gruber Tom Toward Principles for the Design of Ontologies Used for Knowledge Sharing [Artikel] // International Journal of Human-Computer Studies 43 (5-6). - 1995. - S. 907–928.

Grüninger Michael und Fox Mark S. Methodology for the Design and Evaluation of Ontologies [Konferenz] // Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing IJCAI. - Montreal, Canada : [s.n.], 1995.

Grüninger Michael und Fox Mark S. Methodology for the Design and Evaluation of Ontologies. [Buch]. - 1995.

Guarino Nicola und Welty Christopher Evaluating ontological decisions with OntoClean. [Journal] // Communications of the ACM. - 2002. - 2 : Bd. 45. - S. 61.

Hartig Olaf Provenance Information in the Web of Data [Konferenz] // LDOW. - Madrid, Spain : [s.n.], 2009.

Harvard Library Metadata Standards : Digital Projects : Library Technology Services [Online]. - Harvard University Library , 2015. - 25. 04 2015. - http://hul.harvard.edu/ois/digproj/metadata-standards.html.

Heathe Tom und Bizer Christian Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136 [Buch]. - [s.l.] : Morgan & Claypool, 2011.

Hey Tony und Trefethen Anne The Data Deluge: An e-Science Perspective [Journal] // Grid Computing - Making the Global Infrastructure a Reality. - January 2003.

Higgins Sarah The lifecycle of data management [Buchabschnitt] // Managing Research Data / Buchverf. Pryor Graham. - [s.l.] : facet publishing, 2012.

High Level Expert Group on Scientific Data Riding the wave - How Europe can gain from the rising tide of scientific data - Final report of the High Level Expert Group on Scientific Data [Bericht]. - [s.l.] : European Commision, October 2010.

Horrocks Ian, Patel-Schneider Peter F. und Harmelen Frank Van From SHIQ and RDF to OWL: The making of a web ontology language. [Journal] // Web semantics: science, services and agents on the World Wide Web . - 2003. - Bd. 1.1. - S. 7-26.

Hoxha Julia, Rula Anisa und Ell Basil Towards Green Linked Data. [Konferenz] // COLD. - 2011.

International Organization for Standardization ISO 21127:2014 - Information and documentation -- A reference ontology for the interchange of cultural heritage information [Bericht]. - [s.l.] : International Organization for Standardization, 2014.

International Organization for Standardization ISO 2146:2010 Information and documentation — Registry services for libraries and related organizations [Online]. - 2010. - 25. 12 2015. - https://www.iso.org/obp/ui/#iso:std:iso:2146:ed-3:v1:en.

International Organization for Standardization ISO 9000: 2005 - Quality management systems- Fundamentals and vocabulary [Bericht]. - 2005.

Internet Encyclopedia of Philosophy Internet Encyclopedia of Philosophy [Online]. - 2014. - 10 2014. - http://www.iep.utm.edu/prop-log/.

Ireson-Paine Jocelyn Lecture: Rule Based Systems [Online]. - 1996. - 01. 02 2015. - http://www.j-paine.org/students/lectures/lect3/lect3.html.

Jacobson Ivar, Booch Grady und Rumbaugh James The unified software development process [Buch]. - Reading : Addison-wesley, 1999.

Jeschke Sabina [et al.] Networking nanotechnology-resources for scientific education and research with BW-eLabs [Bericht]. - 2009.

Joint Information Systems Committee (UK) SWORD v2 Implementations [Online]. - 2011. - October 2015. - http://swordapp.org/sword-v2/sword-v2-implementations/.

Jörg Brigitte Towards ontological foundations of research information systems [Bericht] : Dissertation / Universität des Saarlandes. - Saarland : Universität des Saarlandes, 2013. - urn:nbn:de:bsz:291-scidok-53947.

Kant Immanuel Critique of pure reason [Buch]. - Cambridge : Cambridge University Press, 1781.

Kindling Maxi und Schirmbacher Peter Die digitale Forschungswelt als Gegenstand der Forschung [Journal] // Information - Wissenschaft & Praxis. - 2013. - 2-3 : Bd. 64. - S. 137-148.

Klump J [et al.] Data publication in the open access initiative [Journal] // Data Science Journal. - 2006. - 5. - S. 79-83.

Klyne Graham und Carroll Jeremy J. Resource description framework (RDF): Concepts and abstract syntax [Bericht]. - [s.l.] : W3C, 2006.

Knowledge Based Systems, Inc. IDEF1X - Data Modeling Method [Online]. - Knowledge Based Systems, Inc, 2014. - 07 2014. - http://www.idef.com/idef1X.htm.

Knublauch Holger [et al.] The Protege OWL Plugin: An Open Development Environment for Semantic Web Applications [Artikel] // The Semantic Web - ISWC 2004. - 2004. - S. 229-243.

Krafft Dean B. [et al.] Vivo: Enabling national networking of scientists. [Journal]. - 2010.

Laningham Scott Doug Lenat on Cyc, a truly semantic Web, and artificial intelligence (AI) [Online]. - IBM, 2008. - 02 2015. - http://www.ibm.com/developerworks/podcast/dwi/cm-int091608txt.html.

Lebo Timothy, Sahoo Satya und McGuinness Deborah PROV-O: The PROV Ontology [Online]. - W3C, 2013. - 02 2015. - http://www.w3.org/TR/prov-o/.

Lebo Timothy, Sahoo Satya und McGuinness Deborah PROV-O: The PROV Ontology - W3C Recommendation 30 April 2013 [Online]. - 30. 04 2013. - 10. 03 2015. - http://www.w3.org/TR/prov-o/.

Lewis Stuart, Castro Pablo De und Jones Richard SWORD: Facilitating deposit scenarios [Journal] // D-Lib Magazine. - 2012. - 1 : Bd. 18.

Library of Congress PREMIS: Preservation Metadata Maintenance Activity [Online]. - Library of Congress, April 2008. - 20. 03 2015. - http://www.loc.gov/standards/premis/.

Liddell Henry George und Scott Robert Perseus Digital Library - A Greek-English Lexicon [Online]. - http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Dmeta%2F, 1920. - 25. 04 2015. - http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0057%3Aentry%3Dmeta%2F.

Lindland Odd Ivar, Sindre Guttorm und Solvberg Arne Understanding quality in conceptual modeling. [Buchabschnitt] // Software, IEEE 11, no. 2. - [s.l.] : IEEE , 1994.

Linux Information Project Metadata definition [Online]. - The Linux Information Project (LINFO), 21. March 2006. - 05. 05 2015. - http://www.linfo.org/metadata.html.

Liu Cixin The Three-Body Problem [Buch]. - Tor Books : [s.n.], 2014.

LODStats LODStats [Online]. - 2015. - 11. September 2015. - http://stats.lod2.eu/.

Lohr Steve For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights [Online]. - New York Times, 17. August 2014. - 11. September 2015. - http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Lord P. [et al.] From Data Deluge to Data Curation [Konferenz] // S. J. Cox, Ed.. - [s.l.] : Proc 3th UK eScience All Hands Meeting, , 2004. - S. 371–375.

Magritte René The treachery of images. - 1929.

Manola Frank und Miller Eric RDF Primer - W3C Recommendation 10 February 2004 [Online]. - W3C , 10. February 2004. - 06 2014. - http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

Masolo Claudio [et al.] WonderWeb Deliverable D18 - Ontology Library (final) [Bericht]. - [s.l.] : ISTC-CNR, 2003.

Matthews Brian [et al.] Using a core scientific metadata model in large-scale facilities [Journal] // International Journal of Digital Curation 5 no. 1. - 2010. - S. 106-118.

McGinnis Jon Avicenna's Naturalized Epistemology and Scientific Method [Buchabschnitt] // Logic, Epistemology, and The Unity of Science. - [s.l.] : Springer Netherlands, 2008.

Montiel-Ponsoda Elena [et al.] Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web [Konferenz]. - 2011.

Moreau Luc [et al.] The open provenance model core specification (v1. 1). [Artikel] // Future Generation Computer Systems 27 No 6. - 2011. - S. 743-756.

Moreau Luc und Missier Paolo PROV-DM: The PROV Data Mode - W3C Recommendation 30 April 2013l [Online]. - 30. 04 2013. - 10. 03 2015. - http://www.w3.org/TR/prov-dm/.

Moreau Luc und Missier Paolo PROV-DM: The PROV Data Model [Online]. - 30. April 2013. - 20. 03 2015. - http://www.w3.org/TR/prov-dm/.

Müller Hans-Michael, Kenny Eimear E. und Sternberg Paul W. Textpresso: an ontology-based information retrieval and extraction system for biological literature [Journal] // PLoS Biol. - 2004. - Bd. 11.

Mylka Antoni [et al.] Nepomuk information element ontology [Bericht]. - 2007.

Mylopoulos John Conceptual Modelling and Telos [Buchabschnitt] // Conceptual Modeling, Databases and Case / Buchverf. Loucopoulos P und Zicari R.. - [s.l.] : Wiley, 1992.

National Institutes of Health National Institutes of Health [Online]. - National Institutes of Health , 26. February 2003. - 2014. - http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

National Research Council On Being a Scientist: A Guide to Responsible Conduct in Research: Third Edition [Buch]. - Washington, DC : The National Academies Press, 2009.

Nature Publishing Group Nature Publishing Group [Online]. - Nature Publishing Group, 2014. - 2014. - http://www.nature.com/authors/policies/availability.html.

Nelson H. James [et al.] A conceptual modeling quality framework. [Buchabschnitt] // Software Quality Journal 20, no. 1. - 2012.

Niles Ian und Pease Adam Mapping WordNet to the SUMO ontology [Artikel] // Proceedings of the IEEE international knowledge engineering conference. - 2003.

Niles Ian und Pease Adam Towards a Standard Upper Ontology [Konferenz] // FOIS. - 2001.

North John Empiricism, Review: Aristotle's [Journal] // Early Science and Medicine . - 2005. - 1 : Bd. 10. - S. 91-97 .

Object Management Group UML 2.2 [Online]. - Object Management Group, Inc., 2009. - 07 2014. - http://www.omg.org/spec/UML/2.2/.

Obrst Leo [et al.] The evaluation of ontologies [Buchabschnitt] // In Semantic Web. - [s.l.] : Springer US, 2007..

Oren Eyal [et al.] Sindice. com: a document-oriented lookup index for open linked data. [Journal] // International Journal of Metadata, Semantics and Ontologies. - 2008. - 1 : Bd. 3. - S. 37-52.

Pease Adam Suggested Upper Merged Ontology (SUMO) [Online]. - Adam Pease, 2002. - 02 2015. - www.ontologyportal.org.

Porzel Robert und Malaka Rainer A task-based approach for ontology evaluation. [Konferenz] // In ECAI Workshop on Ontology Learning and Population. - Valenca, Spain : [s.n.], 2005.

Priem J. [et al.] Altmetrics: A manifesto [Konferenz]. - 2010.

Provenance Challenge Provenance Challenge Wiki [Online]. - Provenance Challenge, 2006. - 03 2015. - http://twiki.ipaw.info/bin/view/Challenge/WebHome.

Prud'hommeaux Eric W3C RDF Validation Service [Online]. - 2006. - 01. 02 2016. - https://www.w3.org/RDF/Validator/.

Ramsay William und Young Sydney On Evaporation and Dissociation.--Part II. A Study of the Thermal Properties of Methyl-Alcohol [Journal] // Philosophical Transactions of the Royal Society of London. - 1886. - Bd. 177.

Razum Matthias und Schwichtenberg Frank Metadatenkonzept für dynamische Daten - BW-eLabs Report [Bericht]. - Karlsruhe : FIZ Karlsruhe, 2012.

Richard Benjamins, V. und Fensel Dieter "Community is Knowledge! in (KA)2 [Journal] // In Proceedings of KAW. - 1998.

Riley Jenn und Becker Devin Seeing Standards: A Visualization of the Metadata Universe [Online]. - Places &amp; Spaces: Mapping Science, 2010. - 25. 04 2015. - http://www.scimaps.org/maps/map/seeing_standards_a_v_130/detail.

Rodriguez-Doncel Victor, Gómez-Pérez Asunción und Mihindukulasooriya Nandana Rights declaration in linked data [Konferenz] // CEUR Workshop Proceedings. - 2013.

Roussey Catherine [et al.] An Introduction to Ontologies and Ontology Engineering. [Buchabschnitt] // Ontologies in Urban Development Projects. - London : Springer, 2011.

Royal Society of London Philosophical Transactions − the world's first science journal [Online]. - Royal Society of London for Improving Natural Knowledge, 2015. - 25. 05 2015. - http://rstl.royalsocietypublishing.org/.

Rudolph Sebastian Foundations of description logics [Buchabschnitt] // Reasoning Web. Semantic Technologies for the Web of Data. - [s.l.] : Springer Berlin Heidelberg, 2011.

Sahoo Satya S. und Sheth Amit P. Provenir ontology: Towards a framework for escience provenance management. [Bericht]. - 2009.

Savage Caroline J. und Andrew J. Vickers Empirical study of data sharing by authors publishing in PLoS journals. [Journal] // PloS one 4. - 2009.

Schirmbacher Peter Aspekte digitaler Informationsversorgung - Forschungsdatenmanagement [Bericht]. - [s.l.] : Präsentationsfollien HU-Berlin (WS 2015-2016), 2015.

Schmidt-Schauß Manfred und Smolka Gert Attributive concept descriptions with complements [Buchabschnitt] // Artificial intelligence 48.1. - 1991.

Schwarte Andreas [et al.] Fedx: Optimization techniques for federated query processing on linked data [Journal] // In The Semantic Web–ISWC. - 2011. - Springer. - S. 601-616.

Shearer Rob, Motik Boris und Horrocks Ian HermiT: A Highly-Efficient OWL Reasoner. [Journal] // OWLED. - 2008. - Bd. 432. - S. 91.

Smith Barry und Ceusters Werner Ontological realism: A methodology for coordinated evolution of scientific ontologies [Artikel] // Applied ontology 5.3-4. - 2010. - S. 139-188.

Soldatova Larisa N., und King Ross D. An ontology of scientific experiments. [Journal] // Journal of the Royal Society Interface. - 2006. - 2011 : Bd. 3. - S. 795-803.

Sprott Haiko und Anderson Gabriella Why publish your negative results? [Online]. - 28. August 2012. - 04 2013. - http://blogs.biomedcentral.com/bmcblog/2012/08/28/why-publish-your-negative-results-2/.

Stonebraker Michael und Hellerstein Joey What goes around comes around [Journal] // Readings in Database Systems 4. - 2005.

Sufi Shoaib und Mathews Brian CCLRC Scientific Metadata Model: Version 2 [Bericht]. - Daresbury, UK : e-Science Centre, CCLRC & Business and Information Technology Department CCLRC, 2004.

Sure York [et al.] OntoEdit: Collaborative ontology development for the semantic web. [Bericht]. - [s.l.] : Springer Berlin Heidelberg, 2002.

Sure York [et al.] On-To-Knowledge: Semantic Web Enabled Knowledge Management [Journal] // Web Intelligence. - 2003. - S. 277-300.

Šváb-Zamazal Ondřej und Svátek Vojtěch Analysing ontological structures through name pattern tracking. [Journal] // In Knowledge Engineering: Practice and Patterns. - 2008. - S. 213-228.

The Dublin Core Metadata Initiative (DCMI) DCMI Metadata Terms - DCMI Usage Board [Online]. - 14. 06 2012. - 10. 03 2015. - http://dublincore.org/documents/dcmi-terms/.

The Royal Society History of the Royal Society [Online]. - 2013. - 2013. - http://royalsociety.org/about-us/history/.

The White House The White House [Online]. - Executive Office of President U.S.A, 22. February 2013. - 2014. - http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Thorndike Lynn Roger Bacon and Experimental Method in the Middle Ages [Journal] // The Philosophical Review . - 1914. - 23 : Bd. 3. - S. 271-298.

Tridgell Andrew und Mackerras Paul The rsync algorithm [Konferenz]. - 1996.

United States Patent and Trademark Office Trademark Electronic Search System (TESS) [Online]. - United States Patent and Trademark Office, 1. September 1998. - 05. 05 2015. - http://tmsearch.uspto.gov/bin/showfield?f=doc&state=4803:u1n5xp.2.24.

Uschold Michael und King Martin Towards a methodology for building ontologies [Buch]. - Edinburgh : University of Edinburgh, 1995.

Uschold Mike Building Ontologies: Towards a Unified Methodology [Journal] // Proceedings of Expert Systems '96, the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems. - 1996.

van der Vlist Eric XML Schema [Buch]. - [s.l.] : O'Reilly Media Inc., 2002.

Villata Serena [et al.] An access control model for linked data [Journal] // On the Move to Meaningful Internet Systems: OTM 2011 Workshops. - 2011. - Springer Berlin Heidelberg. - S. 454-463.

Völker Johanna, Vrandecic Denny und Sure York Automatic Evaluation of Ontologies (AEON [Journal] // Proceedings of the 4th International Semantic Web Conference (ISWC'05). - 2005. - Bd. 3729 . - S. 716-731.

von Humboldt Alexander Reise auf dem Rio Magdalena, durch die Anden und Mexico [Buch]. - 2003. - Margot Faak (ed.) : Bd. 1 : S. 358.

Vrandečić Denny Ontology evaluation [Buch]. - [s.l.] : Springer Berlin Heidelberg, 2010.

W3C Describing Linked Datasets with the VoID Vocabulary - Work in Progress [Online]. - 03. March 2011. - 31. 08 2015. - http://www.w3.org/TR/void/.

W3C Working Group Best Practices for Publishing Linked Data [Online]. - 2014. - 02 2016. - https://www.w3.org/TR/ld-bp/.

Wand Yair und Weber Ron An ontological model of an information system [Buchabschnitt] // Software Engineering, IEEE Transactions on 16, no. 11. - [s.l.] : IEEE , 1990.

Webopedia Linked Data - Webopedia [Online]. - IT Business Edge Network., 2014. - 1. December 2014. - http://www.webopedia.com/TERM/L/Linked_Data.html.

WIPO Berne Convention for the Protection of Literary and Artistic Works [Bericht]. - Bern : WIPO, 1979.

Wolstencroft Katherine [et al.] The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. [Journal] // Nucleic acids research. - 2013. - Bd. gkt328.

Wood David und 3 Round Stones Inc Resource Description Framework [Online]. - W3C, 24. 02 2014. - 20. 08 2015. - http://www.w3.org/TR/rdf11-new/.

Wood John [et al.] Riding the wave: How Europe can gain from the rising tide of scientific data [Bericht] : Report / EC. - 2010.

WordNet About WordNet [Online]. - Princeton University, 2010. - 02 2015. - http://wordnet.princeton.edu.

Workflow Management Coalition Workflow Management Coalition [Online]. - Workflow Management Coalition, 1993. - 03 2015. - http://www.wfmc.org/.

Zaihrayeu I. [et al.] From web directories to ontologies: Natural language processing challenges [Konferenz] // 6th International Semantic Web Conference (ISWC 2007). - 2007..

Zolin Evgeny Description Logic Complexity Navigator [Online]. - 2013. - 01. 02 2016. - http://www.cs.man.ac.uk/~ezolin/dl/.

List of Abbreviations

AI	Artificial Intelligence
BFO	Basic Formal Ontology
BIBO	Bibliographic Ontology
CASRAI	Consortia Advancing Standards in Research Administration
CERIF	Common European Research Information Format
CRIS	Current Research Information System
CWA	Closed-World Assumption
DAML-OIL	DARPA Agent Markup Language – Ontology Inference Layer
DBMS	Database Management Systems
DCAM	DCMI Abstract Model
DCMI	Dublin Core Metadata Initiative
DELOS	Digital Library Reference Model
DOI	Digital Object Identifier
DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
DTD	Document Type Definition
DL	Description Logics
EC	European Commission
EDMS	Electronic Document Management Systems
ESF	European Science Foundation
EXI	Efficient XML Interchange Format
ERM	Entity Relationship Model
FERON	Field-extensible Research Ontology (this work)
FOAF	Friend-of-a-Friend
FRBR	Functional Requirements for Bibliographic Records
ICP	International Cataloguing Principles
ICSU	International Council for Science
IETF	Internet Engineering Task Force
IFLA	International Federation of Library Associations and Institutions
IRI	Internationalized Resource Identifiers
ISBD	International Standard Bibliographic Description
ISO	International Organization for Standardization
JISC	Joint Information Systems Committee
KB	Knowledge Base
KOS	Knowledge Organisation Systems
KR	Knowledge Representation and Reasoning
LD	"The goal of Linked Data is to enable people to share structured data on the Web as easily as they can share documents today" (Bizer, et al., 2007).
LIS	Library and Information Science
LMS	Learning Management System
LOD	Linked Open Data
LOM	Learning Object Metadata
MARC	The MARC formats are standards for the representation and communication of bibliographic and related information in machine-readable form.
MDA	Model-driven Architectures
MDE	Model-driven Engineering
MDD	Model-driven Development
MIS	Management Information Systems
MODS	Metadata Object Description Schema

MOF	Meta-Object-Facility
NLP	Natural Language Processing
NSF	National Science Foundation
OA	Open Access
OAI	Open Access Initiative
OAI-PMH	Open Access Initiative – Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OODBS	Object-Oriented Database System
OPAC	Online Public Access Catalog
ORCID	Open Researcher and Contributor ID
OWL	Web Ontology Language
QA	Question-Answering
RDA	Resource Description and Access
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
REST	Representatioal State Transfer
RIM	Research Information Management
RIS	Research Information System
SKOS	Simple Knowledge Organisation System
SOA	Service-oriented Architectures
SOAP	Simple Object Access Protocol
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SUMO	Suggested Upper Merged Ontology
SUO	Standard Upper Ontology
SW	Semantic Web
URI	Uniform Resource Identifier
UML	Unified Modeling Language
VIVO	An interdisciplinary network of scientists
W3C	World Wide Web Consortium
XML	Extensible Markup Language
XSD	XML Schema Definition Language

List of Figures

List of Tables

Appendix A: Research Questions

Annotation of the result-data with semantic technologies, at their point of creation can produce abundant contextual information to allow reproduction of the investigation and provide a clear view of the process increasing trust in the research processes. In addition, use of semantic technologies to annotate the research-data increases their visibility, usability at the same time address issues such as licensing and access control.

Research Question 1

How can we model the (finite) environment, entities and relations that are part of an investigation process?

Such a model should allow for harvesting of provenance and contextual information containing information on entities such as institution, investigators, study, research and research results.

Research Question 2

How can we use the aforementioned formalization model to simplify the annotation process of research data?

Is it possible to automate the process of data annotation and at what extent?

Research Question 3

How to automate the publishing process of research data in data repositories and still comply with requirements of good scholarly communication practices?

How can this formalization be used to improve the interconnectedness in research activities?

Appendix B: Requirements

Requirement 1: *The model shall contain information on the entities and relations associated to technical aspects of an investigation*

Requirement 2: *The model shall contain information on entities and relations associated to social aspects of an investigation*

Requirement 3: *The model shall support accessibility features to resources; mentioning access to versioning and persistent identifiers to guarantee longevity access*

Requirement 4: *The approach is shown to be feasible for an implementation in an infrastructure*

Requirement 5: *The approach guarantees expressivity and automation*

Requirement 6: *The approach allows for discovery capabilities in a federated environment*

Requirement 7: *The approach will support the concept of interoperability*

Requirement 8: *The approach shall allow for metric information on published data*

Requirement 9: The approach shall guarantee access control support

Requirement 10: *The approach shall provide information on licencing*

Requirement 11: *The formalization shall adhere to linked data (LD) principles for modelling*

Requirement 12: *The formalization shall adhere to LD requirements for publishing*

Requirement 13: *Generalizability of the model and portability of the model to other research disciplines*

Appendix C:
Documentation of Core Ontology of Scientific Investigation

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ns: <http://www.w3.org/2003/06/sw-vocab-status/ns#> .
@prefix nao: <http://www.semanticdesktop.org/ontologies/2007/08/15/nao#> .
@prefix nco: <http://www.semanticdesktop.org/ontologies/2007/03/22/nco#> .
@prefix nfo: <http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#> .
@prefix nie: <http://www.semanticdesktop.org/ontologies/2007/01/19/nie#> .
@prefix nrl: <http://www.semanticdesktop.org/ontologies/2007/08/15/nrl#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix core: <http://purl.org/ontology/olo/core#> .
@prefix prov: <http://www.w3.org/ns/prov-o-20130430#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix vann: <http://purl.org/vocab/vann/> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix prov1: <http://www.w3.org/ns/prov#> .
@prefix terms: <http://purl.org/dc/terms/> .
<http://purl.org/net/cosi#> a owl:Ontology ;
        owl:imports <http://purl.org/ontology/olo/core#> , <http://www.w3.org/ns/prov-o-20130430> ;
        vann:preferredNamespacePrefix "cosi" ;
        rdfs:label "Core Ontology of Scientific Investigation - COSI"@en ;
        owl:versionInfo "0.1"^^xsd:decimal ;
        terms:description "An ontology for the representation of scientific investigation"@en ;
        terms:creator "Armand Brahaj" .
# #    Annotation properties
# http://purl.org/dc/elements/1.1/contributor
dc:contributor a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/coverage
dc:coverage a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/creator
dc:creator a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/date
dc:date a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/description
dc:description a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/format
dc:format a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/identifier
dc:identifier a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/language
dc:language a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/publisher
dc:publisher a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/relation
dc:relation a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/rights
dc:rights a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/source
dc:source a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/subject
dc:subject a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/title
dc:title a owl:AnnotationProperty .
# http://purl.org/dc/elements/1.1/type
dc:type a owl:AnnotationProperty .
# http://purl.org/dc/terms/abstract
terms:abstract a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:description .
# http://purl.org/dc/terms/accessRights
terms:accessRights a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:rights .
# http://purl.org/dc/terms/alternative
terms:alternative a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:title ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/available
terms:available a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/bibliographicCitation
terms:bibliographicCitation a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:identifier ;
        rdfs:range rdfs:Literal ;
        rdfs:domain terms:BibliographicResource .
# http://purl.org/dc/terms/conformsTo
terms:conformsTo a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/contributor
```

```
terms:contributor a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:contributor .
# http://purl.org/dc/terms/coverage
terms:coverage a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:coverage .
# http://purl.org/dc/terms/created
terms:created a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/creator
terms:creator a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:creator .
# http://purl.org/dc/terms/date
terms:date a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/dateAccepted
terms:dateAccepted a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/dateCopyrighted
terms:dateCopyrighted a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/dateSubmitted
terms:dateSubmitted a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/description
terms:description a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:description ;
        rdfs:domain :Onymous .
# http://purl.org/dc/terms/extent
terms:extent a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:format .
# http://purl.org/dc/terms/format
terms:format a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:format .
# http://purl.org/dc/terms/hasFormat
terms:hasFormat a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/hasPart
terms:hasPart a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/hasVersion
terms:hasVersion a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/identifier
terms:identifier a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:identifier ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/isFormatOf
terms:isFormatOf a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/isPartOf
terms:isPartOf a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/isReferencedBy
terms:isReferencedBy a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/isReplacedBy
terms:isReplacedBy a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/isRequiredBy
terms:isRequiredBy a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/isVersionOf
terms:isVersionOf a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/issued
terms:issued a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/language
terms:language a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:language .
# http://purl.org/dc/terms/license
terms:license a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:rights .
# http://purl.org/dc/terms/medium
terms:medium a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:format .
# http://purl.org/dc/terms/modified
terms:modified a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/publisher
terms:publisher a owl:AnnotationProperty ;
```

```
        rdfs:subPropertyOf dc:publisher .
# http://purl.org/dc/terms/references
terms:references a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/relation
terms:relation a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation .
# http://purl.org/dc/terms/replaces
terms:replaces a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/requires
terms:requires a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:relation , terms:relation .
# http://purl.org/dc/terms/rights
terms:rights a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:rights .
# http://purl.org/dc/terms/source
terms:source a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:source , terms:relation .
# http://purl.org/dc/terms/spatial
terms:spatial a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:coverage .
# http://purl.org/dc/terms/subject
terms:subject a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:subject .
# http://purl.org/dc/terms/tableOfContents
terms:tableOfContents a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:description .
# http://purl.org/dc/terms/temporal
terms:temporal a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:coverage .
# http://purl.org/dc/terms/title
terms:title a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:title ;
        rdfs:range rdfs:Literal .
# http://purl.org/dc/terms/type
terms:type a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:type ;
        rdfs:range rdfs:Class .
# http://purl.org/dc/terms/valid
terms:valid a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range rdfs:Literal .
# http://purl.org/net/cosi#hasDescription
:hasDescription a owl:AnnotationProperty ;
        rdfs:domain :Onymous .
# http://purl.org/net/cosi#hasIdentifier
:hasIdentifier a owl:AnnotationProperty ;
        rdfs:range :Onymous .
# http://purl.org/net/cosi#hasTitle
:hasTitle a owl:AnnotationProperty ;
        rdfs:domain :Onymous .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#byteSize
nie:byteSize a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#created
nie:created a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#hasLogicalPart
nie:hasLogicalPart a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#hasPart
nie:hasPart a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#identifier
nie:identifier a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#isPartOf
nie:isPartOf a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#lastModified
nie:lastModified a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#links
nie:links a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#url
nie:url a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nco#creator
nco:creator a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#belongsToContainer
nfo:belongsToContainer a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:isPartOf ;
        rdfs:range nfo:DataContainer .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#bookmarks
nfo:bookmarks a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:links ;
        rdfs:range nie:DataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#containsBookmark
nfo:containsBookmark a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:hasLogicalPart .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#containsBookmarkFolder
nfo:containsBookmarkFolder a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:hasLogicalPart .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#containsPlacemark
nfo:containsPlacemark a owl:AnnotationProperty ;
```

```
        rdfs:subPropertyOf nie:hasLogicalPart .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileCreated
nfo:fileCreated a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:created ;
        rdfs:range xsd:dateTime ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileLastAccessed
nfo:fileLastAccessed a owl:AnnotationProperty ;
        rdfs:subPropertyOf dc:date ;
        rdfs:range xsd:dateTime ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileLastModified
nfo:fileLastModified a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:lastModified ;
        rdfs:range xsd:dateTime ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileName
nfo:fileName a owl:AnnotationProperty ;
        rdfs:subPropertyOf nao:prefLabel ;
        rdfs:range xsd:string ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileSize
nfo:fileSize a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:byteSize ;
        rdfs:range xsd:integer ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileUrl
nfo:fileUrl a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:url ;
        rdfs:range rdfs:Resource .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#foundry
nfo:foundry a owl:AnnotationProperty ;
        rdfs:subPropertyOf nco:creator .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hasMediaStream
nfo:hasMediaStream a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:hasPart ;
        rdfs:range nie:DataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#uuid
nfo:uuid a owl:AnnotationProperty ;
        rdfs:subPropertyOf nie:identifier ;
        rdfs:range xsd:string ;
        rdfs:domain nfo:Filesystem .
# http://www.semanticdesktop.org/ontologies/2007/08/15/nao#deprecated
nao:deprecated a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/08/15/nao#prefLabel
nao:prefLabel a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/08/15/nao#userVisible
nao:userVisible a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/08/15/nrl#cardinality
nrl:cardinality a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/08/15/nrl#inverseProperty
nrl:inverseProperty a owl:AnnotationProperty .
# http://www.semanticdesktop.org/ontologies/2007/08/15/nrl#maxCardinality
nrl:maxCardinality a owl:AnnotationProperty .
# http://www.w3.org/2000/01/rdf-schema#comment
rdfs:comment a owl:AnnotationProperty .
# http://www.w3.org/2000/01/rdf-schema#isDefinedBy
rdfs:isDefinedBy a owl:AnnotationProperty .
# http://www.w3.org/2000/01/rdf-schema#label
rdfs:label a owl:AnnotationProperty .
# http://www.w3.org/2000/01/rdf-schema#seeAlso
rdfs:seeAlso a owl:AnnotationProperty .
# http://www.w3.org/2002/07/owl#versionInfo
owl:versionInfo a owl:AnnotationProperty .
# http://www.w3.org/2004/02/skos/core#note
skos:note a owl:AnnotationProperty .
# http://www.w3.org/ns/prov#definition
prov1:definition a owl:AnnotationProperty .
# #    Datatypes
# http://purl.org/dc/terms/Box
terms:Box a rdfs:Datatype .
# http://purl.org/dc/terms/ISO3166
terms:ISO3166 a rdfs:Datatype .
# http://purl.org/dc/terms/ISO639-2
terms:ISO639-2 a rdfs:Datatype .
# http://purl.org/dc/terms/ISO639-3
terms:ISO639-3 a rdfs:Datatype .
# http://purl.org/dc/terms/Period
terms:Period a rdfs:Datatype .
# http://purl.org/dc/terms/Point
terms:Point a rdfs:Datatype .
# http://purl.org/dc/terms/RFC1766
terms:RFC1766 a rdfs:Datatype .
# http://purl.org/dc/terms/RFC3066
terms:RFC3066 a rdfs:Datatype .
# http://purl.org/dc/terms/RFC4646
terms:RFC4646 a rdfs:Datatype .
# http://purl.org/dc/terms/RFC5646
```

```
terms:RFC5646 a rdfs:Datatype .
# http://purl.org/dc/terms/URI
terms:URI a rdfs:Datatype .
# http://purl.org/dc/terms/W3CDTF
terms:W3CDTF a rdfs:Datatype .
# http://www.w3.org/1999/02/22-rdf-syntax-ns#PlainLiteral
rdf:PlainLiteral a rdfs:Datatype .
# http://www.w3.org/2000/01/rdf-schema#Literal
rdfs:Literal a rdfs:Datatype .
# http://www.w3.org/2001/XMLSchema#date
xsd:date a rdfs:Datatype .
# http://www.w3.org/2001/XMLSchema#dateTime
xsd:dateTime a rdfs:Datatype .
# http://www.w3.org/2001/XMLSchema#duration
xsd:duration a rdfs:Datatype .
# #    Object Properties
# http://purl.org/dc/terms/accessRights
terms:accessRights a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:rights .
# http://purl.org/dc/terms/accrualMethod
terms:accrualMethod a owl:ObjectProperty .
# http://purl.org/dc/terms/accrualPeriodicity
terms:accrualPeriodicity a owl:ObjectProperty .
# http://purl.org/dc/terms/accrualPolicy
terms:accrualPolicy a owl:ObjectProperty ;
        rdfs:range terms:Policy .
# http://purl.org/dc/terms/audience
terms:audience a owl:ObjectProperty .
# http://purl.org/dc/terms/conformsTo
terms:conformsTo a owl:ObjectProperty .
# http://purl.org/dc/terms/contributor
terms:contributor a owl:ObjectProperty .
# http://purl.org/dc/terms/coverage
terms:coverage a owl:ObjectProperty .
# http://purl.org/dc/terms/creator
terms:creator a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:contributor .
# http://purl.org/dc/terms/educationLevel
terms:educationLevel a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:audience .
# http://purl.org/dc/terms/extent
terms:extent a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:format .
# http://purl.org/dc/terms/format
terms:format a owl:ObjectProperty .
# http://purl.org/dc/terms/instructionalMethod
terms:instructionalMethod a owl:ObjectProperty .
# http://purl.org/dc/terms/language
terms:language a owl:ObjectProperty .
# http://purl.org/dc/terms/license
terms:license a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:rights .
# http://purl.org/dc/terms/mediator
terms:mediator a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:audience .
# http://purl.org/dc/terms/medium
terms:medium a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:format .
# http://purl.org/dc/terms/provenance
terms:provenance a owl:ObjectProperty .
# http://purl.org/dc/terms/publisher
terms:publisher a owl:ObjectProperty .
# http://purl.org/dc/terms/rights
terms:rights a owl:ObjectProperty .
# http://purl.org/dc/terms/rightsHolder
terms:rightsHolder a owl:ObjectProperty .
# http://purl.org/dc/terms/spatial
terms:spatial a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:coverage .
# http://purl.org/dc/terms/temporal
terms:temporal a owl:ObjectProperty ;
        rdfs:subPropertyOf terms:coverage .
# http://purl.org/dc/terms/type
terms:type a owl:ObjectProperty .
# http://purl.org/net/cosi#acceptedInput
:acceptedInput a owl:ObjectProperty ;
        rdfs:domain :Instrument ;
        rdfs:comment "Provides information on the accepted format or composition of the input information"@en ;
        rdfs:label "acceptedInput"@en .
# http://purl.org/net/cosi#belongsToProgramme
:belongsToProgramme a owl:ObjectProperty ;
        rdfs:domain :Study ;
        rdfs:range :Programme ;
        rdfs:label "belongsToProgramme"@en .
# http://purl.org/net/cosi#belongsToResearchFacility
:belongsToResearchFacility a owl:ObjectProperty ;
        rdfs:domain :Instrument ;
        rdfs:range :ResearchFacility ;
        rdfs:comment "Relates an instrument with a location"@en ;
        rdfs:label "belongsToResearchFacility"@en .
# http://purl.org/net/cosi#belongsToStudy
:belongsToStudy a owl:ObjectProperty ;
        owl:inverseOf :hasInvestigation ;
        rdfs:domain :Investigation ;
        rdfs:label "belongsToStudy"@en .
# http://purl.org/net/cosi#dataOutput
:dataOutput a owl:ObjectProperty ;
        rdfs:domain :DigitalInstrument ;
        rdfs:comment "Provides information on the result output."@en ;
        rdfs:label "dataOutput"@en .
# http://purl.org/net/cosi#derivesFromResult
:derivesFromResult a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :Result .
# http://purl.org/net/cosi#generatedResultSet
:generatedResultSet a owl:ObjectProperty ;
        owl:inverseOf :isGeneratedByInstrument ;
        rdfs:domain :Instrument ;
        rdfs:range :ResultSet .
# http://purl.org/net/cosi#hasActivity
:hasActivity a owl:ObjectProperty ;
        rdfs:domain :Process ;
        rdfs:range prov1:Activity .
# http://purl.org/net/cosi#hasCalibration
:hasCalibration a owl:ObjectProperty ;
        rdfs:domain :Instrument ;
        rdfs:comment "Relates an instrument with a specific of calibrations"@en ;
        rdfs:label "hasCalibration"@en .
# http://purl.org/net/cosi#hasConclusion
:hasConclusion a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :Conclusion ;
        rdfs:label "hasConclusion"@en .
# http://purl.org/net/cosi#hasDigitalModel
:hasDigitalModel a owl:ObjectProperty ;
        rdfs:domain :Simulation ;
        rdfs:label "hasDigitalModel"@en .
# http://purl.org/net/cosi#hasExperimentProcedure
:hasExperimentProcedure a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasProcedure ;
        rdfs:domain :Experiment ;
        rdfs:range :ExperimentProcedure .
# http://purl.org/net/cosi#hasFieldOfStudy
:hasFieldOfStudy a owl:ObjectProperty ;
        rdfs:domain :Investigation ;
        rdfs:range :AcademicDiscipline .
# http://purl.org/net/cosi#hasHypothesis
:hasHypothesis a owl:ObjectProperty ;
        rdfs:domain :Investigation ;
        rdfs:range :Hypothesis .
# http://purl.org/net/cosi#hasHypothesisAcceptance
:hasHypothesisAcceptance a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :HypothesisAcceptance ;
        rdfs:label "hasHypothesisAcceptance"@en .
# http://purl.org/net/cosi#hasHypothesisRejection
:hasHypothesisRejection a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :HypothesisRejection .
# http://purl.org/net/cosi#hasHypothesisStatement
:hasHypothesisStatement a owl:ObjectProperty ;
        rdfs:domain :Hypothesis ;
        rdfs:range :HypothesisStatement .
# http://purl.org/net/cosi#hasIdentifier
:hasIdentifier a owl:ObjectProperty ;
        rdfs:domain :Preservation , nie:DataObject , nie:InformationElement .
# http://purl.org/net/cosi#hasInput
:hasInput a owl:ObjectProperty ;
        rdfs:domain :Process ;
        rdfs:range :Input ;
        rdfs:label "hasInput"@en .
# http://purl.org/net/cosi#hasInstrument
:hasInstrument a owl:ObjectProperty ;
        rdfs:domain :Experiment , :Measurement , :Rig ;
        rdfs:range :Instrument ;
        rdfs:label "hasInstrument"@en .
# http://purl.org/net/cosi#hasInvestigation
:hasInvestigation a owl:ObjectProperty ;
        rdfs:domain :Study ;
        rdfs:range :Investigation .
# http://purl.org/net/cosi#hasInvestigationProcedure
:hasInvestigationProcedure a owl:ObjectProperty ;
        rdfs:domain :Investigation ;
        rdfs:range :InvestigationProcedure .
# http://purl.org/net/cosi#hasInvestigator
```

```
:hasInvestigator a owl:ObjectProperty ;
        owl:inverseOf :isInvestigatorOf ;
        rdfs:domain :Investigation , :Study ;
        rdfs:range :Investigator .
# http://purl.org/net/cosi#hasLicenceInformation
:hasLicenceInformation a owl:ObjectProperty ;
        rdfs:domain :ResultSet , :Study , nie:InformationElement ;
        rdfs:range :LicenceInformation ;
        rdfs:comment "Relates a Study with one ore more LegalNotes"@en ;
        rdfs:label "hasLicenceInformation"@en .
# http://purl.org/net/cosi#hasMeasurementProcedure
:hasMeasurementProcedure a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasProcedure .
# http://purl.org/net/cosi#hasModel
:hasModel a owl:ObjectProperty ;
        rdfs:domain :Instrument ;
        rdfs:comment "Relates an instrument to a model number obtained by the vendor"@en ;
        rdfs:label "hasModel"@en .
# http://purl.org/net/cosi#hasMonitoredFolder
:hasMonitoredFolder a owl:ObjectProperty ;
        rdfs:domain :DigitalInstrument ;
        rdfs:range nfo:FileDataObject ;
        rdfs:comment "Provides information on a path where the information is saved from the instrument"@en
;
        rdfs:label "hasMonitoredFolder"@en .
# http://purl.org/net/cosi#hasMoticationPublication
:hasMoticationPublication a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasPublication ;
        rdfs:domain :Investigation , :Study ;
        rdfs:range :Motivation_Publication .
# http://purl.org/net/cosi#hasObservationProcedure
:hasObservationProcedure a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasProcedure ;
        rdfs:domain :Observation ;
        rdfs:range :ObservationProcedure .
# http://purl.org/net/cosi#hasObserver
:hasObserver a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasInvestigator ;
        owl:inverseOf :isObserverOf ;
        rdfs:domain :Observation , :ObservationProcedure ;
        rdfs:range :Observer .
# http://purl.org/net/cosi#hasOperatingSystem
:hasOperatingSystem a owl:ObjectProperty ;
        rdfs:domain :DigitalInstrument ;
        rdfs:comment "Provides information on an operating system for the instrument"@en ;
        rdfs:label "hasOperatingSystem"@en .
# http://purl.org/net/cosi#hasOrderederProcedure
:hasOrderederProcedure a owl:ObjectProperty ;
        rdfs:subPropertyOf core:ordered_list ;
        rdfs:domain :ExperimentProcedure ;
        rdfs:range core:OrderedList .
# http://purl.org/net/cosi#hasOutput
:hasOutput a owl:ObjectProperty ;
        rdfs:domain :Process ;
        rdfs:range :Output ;
        rdfs:label "hasOutput"@en .
# http://purl.org/net/cosi#hasParameter
:hasParameter a owl:ObjectProperty ;
        rdfs:domain :DigitalInstrument , :Experiment , :Measurement ;
        rdfs:range :Parameter ;
        rdfs:label "hasParameter"@en .
# http://purl.org/net/cosi#hasPersistendIdentifier
:hasPersistendIdentifier a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasIdentifier ;
        rdfs:domain :Preservation ;
        rdfs:range :PersistendIdentifier .
# http://purl.org/net/cosi#hasPhysicalPath
:hasPhysicalPath a owl:ObjectProperty ;
        rdfs:domain :Preservation ;
        rdfs:range nie:DataObject .
# http://purl.org/net/cosi#hasPreviousInvestigation
:hasPreviousInvestigation a owl:ObjectProperty ;
        rdfs:domain :Investigation , :Study ;
        rdfs:range :Investigation , :Study .
# http://purl.org/net/cosi#hasPriorPreservationVersion
:hasPriorPreservationVersion a owl:ObjectProperty ;
        rdfs:domain :Preservation ;
        rdfs:range :Preservation .
# http://purl.org/net/cosi#hasProcedure
:hasProcedure a owl:ObjectProperty ;
        rdfs:domain :Investigation , :Procedure ;
        rdfs:range :Procedure .
# http://purl.org/net/cosi#hasPublication
:hasPublication a owl:ObjectProperty ;
        rdfs:domain :Investigation , :Study ;
        rdfs:range :Publication , :ScholarCommunication , nie:InformationElement .
# http://purl.org/net/cosi#hasRelatedMaterial
:hasRelatedMaterial a owl:ObjectProperty ;
```

```
        rdfs:domain :Study ;
        rdfs:range nie:InformationElement ;
        rdfs:comment """Relates a Study with one ore more Document types
Might be a follow up publication, a technical report etc"""@en ;
        rdfs:label "hasRelatedMaterial"@en .
# http://purl.org/net/cosi#hasResearchFacility
:hasResearchFacility a owl:ObjectProperty ;
        rdfs:domain :Experiment ;
        rdfs:range :ResearchFacility .
# http://purl.org/net/cosi#hasResult
:hasResult a owl:ObjectProperty ;
        rdfs:subPropertyOf owl:topObjectProperty ;
        rdfs:domain :Investigation , :Publication , :Study ;
        rdfs:range :Result ;
        rdfs:label "hasResult"@en .
# http://purl.org/net/cosi#hasResultConclusion
:hasResultConclusion a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :S6_Conclusion .
# http://purl.org/net/cosi#hasResultError
:hasResultError a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :ResultError .
# http://purl.org/net/cosi#hasResultPublication
:hasResultPublication a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasPublication ;
        rdfs:domain :Investigation ;
        rdfs:range :Result_Publication .
# http://purl.org/net/cosi#hasResultSet
:hasResultSet a owl:ObjectProperty ;
        rdfs:domain :Result ;
        rdfs:range :ResultSet , nie:DataObject ;
        rdfs:comment "The relation to a result to the final dataset, or digital objects containing the
outcome"@en ;
        rdfs:label "hasResultSet"@en .
# http://purl.org/net/cosi#hasRig
:hasRig a owl:ObjectProperty ;
        rdfs:domain :Experiment , :ExperimentProcedure ;
        rdfs:range :Rig ;
        rdfs:label "hasRig"@en .
# http://purl.org/net/cosi#hasRightsDocument
:hasRightsDocument a owl:ObjectProperty ;
        rdfs:domain :RightsStatement ;
        rdfs:range :RightsDocumentation .
# http://purl.org/net/cosi#hasRightsStatement
:hasRightsStatement a owl:ObjectProperty ;
        rdfs:domain :Motivation_Publication , :Publication , :ResultSet , :Result_Publication ,
:RightsDocumentation , nie:DataObject , nie:InformationElement , nfo:ArchiveItem , nfo:Attachment ,
nfo:DataContainer , nfo:DeletedResource , nfo:EmbeddedFileDataObject , nfo:FileDataObject , nfo:Folder ,
nfo:HtmlDocument , nfo:LocalFileDataObject , nfo:PaginatedTextDocument , nfo:PlainTextDocument ,
nfo:Presentation , nfo:RemoteDataObject , nfo:SourceCode , nfo:Spreadsheet , nfo:TextDocument , nfo:Website
, prov1:Entity ;
        rdfs:range :RightsStatement .
# http://purl.org/net/cosi#hasSimulationProcedure
:hasSimulationProcedure a owl:ObjectProperty ;
        rdfs:subPropertyOf :hasProcedure .
# http://purl.org/net/cosi#hasStatus
:hasStatus a owl:ObjectProperty ;
        rdfs:domain :Study ;
        rdfs:comment "Relates the study to a literal or enumerated status"@en ;
        rdfs:label "hasStatus"@en .
# http://purl.org/net/cosi#hasStudy
:hasStudy a owl:ObjectProperty ;
        owl:inverseOf :runByInstitution ;
        rdfs:domain :Programme ;
        rdfs:range :Study .
# http://purl.org/net/cosi#hasThroughput
:hasThroughput a owl:ObjectProperty ;
        rdfs:domain :DigitalInstrument ;
        rdfs:comment "Provides information on the sum of the data rates that are delivered by the
instrument"@en ;
        rdfs:label "hasThroughput"@en .
# http://purl.org/net/cosi#hasTopic
:hasTopic a owl:ObjectProperty ;
        rdfs:domain :Study ;
        rdfs:range :Topic ;
        rdfs:label "hasTopic"@en .
# http://purl.org/net/cosi#hasVendor
:hasVendor a owl:ObjectProperty ;
        rdfs:domain :Instrument ;
        rdfs:comment "Relates an instrument to a vendor"@en ;
        rdfs:label "hasVendor"@en .
# http://purl.org/net/cosi#isGeneratedByInstrument
:isGeneratedByInstrument a owl:ObjectProperty .
# http://purl.org/net/cosi#isInvestigatorOf
:isInvestigatorOf a owl:ObjectProperty .
# http://purl.org/net/cosi#isObserverOf
:isObserverOf a owl:ObjectProperty ;
```

```
        rdfs:domain :Observer ;
        rdfs:range :Observation ;
        rdfs:label "isObserverOf"@en .
# http://purl.org/net/cosi#isPartOf
:isPartOf a owl:ObjectProperty ;
        rdfs:domain :Instrument ;
        rdfs:range :Rig ;
        rdfs:comment "Relates an Instrument with a Rig as part of system, part of a platform or
pipeline"@en ;
        rdfs:label "isPartOf"@en .
# http://purl.org/net/cosi#isResearchFacilityOfExperiment
:isResearchFacilityOfExperiment a owl:ObjectProperty ;
        rdfs:domain :ResearchFacility ;
        rdfs:range :Experiment .
# http://purl.org/net/cosi#returnsDataFormat
:returnsDataFormat a owl:ObjectProperty ;
        rdfs:domain :DigitalInstrument ;
        rdfs:comment "Provides information on the data format which the instrument will store the data"@en
;
        rdfs:label "returnsDataFormat"@en .
# http://purl.org/net/cosi#runByInstitution
:runByInstitution a owl:ObjectProperty ;
        rdfs:domain :Investigation , :Study ;
        rdfs:range :Institution .
# http://purl.org/net/cosi#runsOnComputer
:runsOnComputer a owl:ObjectProperty ;
        rdfs:domain :Simulation ;
        rdfs:range :Computer .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#belongsToContainer
nfo:belongsToContainer a owl:ObjectProperty ;
        rdfs:domain nie:DataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#bookmarks
nfo:bookmarks a owl:ObjectProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#compressionType
nfo:compressionType a owl:ObjectProperty ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The type of the compression. Values include, 'lossy' and 'lossless'." ;
        rdfs:label "compressionType" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#containsBookmark
nfo:containsBookmark a owl:ObjectProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#containsBookmarkFolder
nfo:containsBookmarkFolder a owl:ObjectProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#containsPlacemark
nfo:containsPlacemark a owl:ObjectProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#depiction
nfo:depiction a owl:ObjectProperty ;
        nrl:inverseProperty nfo:depicts ;
        rdfs:comment "Relates an information element to an image which depicts said element." ;
        rdfs:label "depiction" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#depicts
nfo:depicts a owl:ObjectProperty ;
        nrl:inverseProperty nfo:depiction ;
        rdfs:comment "Relates an image to the information elements it depicts." ;
        rdfs:label "depicts" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#encryptionStatus
nfo:encryptionStatus a owl:ObjectProperty ;
        rdfs:domain nie:InformationElement ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The status of the encryption of the InformationElement." ;
        rdfs:label "encryptionStatus" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileOwner
nfo:fileOwner a owl:ObjectProperty ;
        rdfs:domain nfo:FileDataObject ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The owner of the file as defined by the file system access rights feature." ;
        rdfs:label "fileOwner" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileUrl
nfo:fileUrl a owl:ObjectProperty ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#foundry
nfo:foundry a owl:ObjectProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hasHash
nfo:hasHash a owl:ObjectProperty ;
        rdfs:domain nfo:FileDataObject .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hasMediaFileListEntry
nfo:hasMediaFileListEntry a owl:ObjectProperty ;
        rdfs:comment "This property is intended to point to an RDF list of MediaFiles." ;
        rdfs:label "hasMediaFileListEntry" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hasMediaStream
nfo:hasMediaStream a owl:ObjectProperty .
# http://www.w3.org/ns/prov#generated
prov1:generated a owl:ObjectProperty .
# http://www.w3.org/ns/prov#used
prov1:used a owl:ObjectProperty .
# #    Data properties
# http://purl.org/dc/terms/abstract
terms:abstract a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:description .
```

```
# http://purl.org/dc/terms/alternative
terms:alternative a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:title .
# http://purl.org/dc/terms/available
terms:available a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/bibliographicCitation
terms:bibliographicCitation a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:identifier .
# http://purl.org/dc/terms/created
terms:created a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/date
terms:date a owl:DatatypeProperty .
# http://purl.org/dc/terms/dateAccepted
terms:dateAccepted a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/dateCopyrighted
terms:dateCopyrighted a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/dateSubmitted
terms:dateSubmitted a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/description
terms:description a owl:DatatypeProperty .
# http://purl.org/dc/terms/identifier
terms:identifier a owl:DatatypeProperty .
# http://purl.org/dc/terms/issued
terms:issued a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/modified
terms:modified a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/dc/terms/tableOfContents
terms:tableOfContents a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:description .
# http://purl.org/dc/terms/title
terms:title a owl:DatatypeProperty .
# http://purl.org/dc/terms/valid
terms:valid a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:date .
# http://purl.org/net/cosi#hasDescription
:hasDescription a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:description .
# http://purl.org/net/cosi#hasDigestAlgorithm
:hasDigestAlgorithm a owl:DatatypeProperty ;
        rdfs:domain :Preservation ;
        prov1:definition "Ex: MD5 hash code" .
# http://purl.org/net/cosi#hasEndOfLife
:hasEndOfLife a owl:DatatypeProperty ;
        rdfs:domain :Preservation ;
        prov1:definition "Foreseen End Of Life for the Preservation Action." .
# http://purl.org/net/cosi#hasFirstPublicDate
:hasFirstPublicDate a owl:DatatypeProperty ;
        rdfs:domain :Preservation .
# http://purl.org/net/cosi#hasHandler
:hasHandler a owl:DatatypeProperty .
# http://purl.org/net/cosi#hasHypothesis
:hasHypothesis a owl:DatatypeProperty ;
        rdfs:range rdfs:Literal .
# http://purl.org/net/cosi#hasIdentifier
:hasIdentifier a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:identifier .
# http://purl.org/net/cosi#hasIdentifierValue
:hasIdentifierValue a owl:DatatypeProperty .
# http://purl.org/net/cosi#hasIngestDate
:hasIngestDate a owl:DatatypeProperty ;
        rdfs:domain :Preservation .
# http://purl.org/net/cosi#hasPreservationStatus
:hasPreservationStatus a owl:DatatypeProperty ;
        prov1:definition "A status for the preservation that indicates availability of the resources. Eg.
Public, Withdrawn, Pending Delete, Public" .
# http://purl.org/net/cosi#hasTitle
:hasTitle a owl:DatatypeProperty ;
        rdfs:subPropertyOf terms:title .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#aspectRatio
nfo:aspectRatio a owl:DatatypeProperty ;
        rdfs:range xsd:float ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Visual content aspect ratio. (Width divided by Height)" ;
        rdfs:label "aspectRatio" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#averageBitrate
nfo:averageBitrate a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:rate ;
        rdfs:range xsd:float ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The average overall bitrate of a media container. (i.e. the size of the piece of
media in bits, divided by it's duration expressed in seconds)." ;
```

```
        rdfs:label "averageBitrate" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#bitDepth
nfo:bitDepth a owl:DatatypeProperty ;
        rdfs:range rdfs:Literal ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "A common superproperty for all properties signifying the amount of bits for an atomic
unit of data. Examples of subproperties may include bitsPerSample and bitsPerPixel" ;
        rdfs:label "bitDepth" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#bitrateType
nfo:bitrateType a owl:DatatypeProperty ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The type of the bitrate. Examples may include CBR and VBR." ;
        rdfs:label "bitrateType" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#bitsPerSample
nfo:bitsPerSample a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:bitDepth ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Amount of bits in each audio sample." ;
        rdfs:label "bitsPerSample" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#channels
nfo:channels a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Number of channels. This property is to be used directly if no detailed information
is necessary. Otherwise use more detailed subproperties." ;
        rdfs:label "channels" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#characterCount
nfo:characterCount a owl:DatatypeProperty ;
        rdfs:domain nfo:TextDocument ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of characters in the document." ;
        rdfs:label "characterCount" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#characterPosition
nfo:characterPosition a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Character position of the bookmark." ;
        rdfs:label "characterPosition" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#codec
nfo:codec a owl:DatatypeProperty ;
        rdfs:range rdfs:Literal ;
        rdfs:comment "The name of the codec necessary to decode a piece of media." ;
        rdfs:label "codec" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#colorCount
nfo:colorCount a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The number of colors used/available in a raster image." ;
        rdfs:label "color count" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#colorDepth
nfo:colorDepth a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:bitDepth ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Amount of bits used to express the color of each pixel." ;
        rdfs:label "colorDepth" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#commentCharacterCount
nfo:commentCharacterCount a owl:DatatypeProperty ;
        rdfs:domain nfo:SourceCode ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of character in comments i.e. characters ignored by the
compiler/interpreter." ;
        rdfs:label "commentCharacterCount" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#count
nfo:count a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        rdfs:comment "A common superproperty for all properties signifying the amount of atomic media data
units. Examples of subproperties may include sampleCount and frameCount." ;
        rdfs:label "count" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#definesClass
nfo:definesClass a owl:DatatypeProperty ;
        rdfs:domain nfo:SourceCode ;
        rdfs:range xsd:string ;
        rdfs:comment "Name of a class defined in the source code file." ;
        rdfs:label "definesClass" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#definesFunction
nfo:definesFunction a owl:DatatypeProperty ;
        rdfs:domain nfo:SourceCode ;
        rdfs:range xsd:string ;
        rdfs:comment "A name of a function/method defined in the given source code file." ;
        rdfs:label "definesFunction" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#definesGlobalVariable
nfo:definesGlobalVariable a owl:DatatypeProperty ;
        rdfs:domain nfo:SourceCode ;
        rdfs:range xsd:string ;
        rdfs:comment "Name of a global variable defined within the source code file." ;
        rdfs:label "definesGlobalVariable" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#deletionDate
nfo:deletionDate a owl:DatatypeProperty ;
        rdfs:domain nfo:DeletedResource ;
        rdfs:range xsd:dateTime ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The date and time of the deletion." ;
        rdfs:label "deletionDate" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#duration
nfo:duration a owl:DatatypeProperty ;
        rdfs:range xsd:duration ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Duration of a media piece." ;
        rdfs:label "duration" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#encoding
nfo:encoding a owl:DatatypeProperty ;
        rdfs:domain nfo:EmbeddedFileDataObject ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The encoding used for the Embedded File. Examples might include BASE64 or UUEncode" ;
        rdfs:label "encoding" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileCreated
nfo:fileCreated a owl:DatatypeProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileLastAccessed
nfo:fileLastAccessed a owl:DatatypeProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileLastModified
nfo:fileLastModified a owl:DatatypeProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileName
nfo:fileName a owl:DatatypeProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fileSize
nfo:fileSize a owl:DatatypeProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#filesystemType
nfo:filesystemType a owl:DatatypeProperty ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Type of filesystem such as ext3 and ntfs." ;
        rdfs:label "filesystemType" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#fontFamily
nfo:fontFamily a owl:DatatypeProperty ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The name of the font family." ;
        rdfs:label "fontFamily" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#frameCount
nfo:frameCount a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:count ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of frames in a video sequence." ;
        rdfs:label "frameCount" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#frameRate
nfo:frameRate a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:rate ;
        rdfs:range xsd:float ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Amount of video frames per second." ;
        rdfs:label "frameRate" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#freeSpace
nfo:freeSpace a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Unoccupied storage space of the filesystem." ;
        rdfs:label "freeSpace" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#frontChannels
nfo:frontChannels a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:channels ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Number of front channels." ;
        rdfs:label "frontChannels" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hashAlgorithm
nfo:hashAlgorithm a owl:DatatypeProperty ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Name of the algorithm used to compute the hash value. Examples might include CRC32,
MD5, SHA, TTH etc." ;
        rdfs:label "hashAlgorithm" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hashValue
nfo:hashValue a owl:DatatypeProperty ;
        rdfs:range xsd:string ;
        nrl:cardinality "1"^^xsd:integer ;
        rdfs:comment "The actual value of the hash." ;
        rdfs:label "hashValue" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#height
nfo:height a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Visual content height in pixels." ;
        rdfs:label "height" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#horizontalResolution
nfo:horizontalResolution a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Horizontal resolution of an image (if printed). Expressed in DPI." ;
        rdfs:label "horizontalResolution" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#interlaceMode
nfo:interlaceMode a owl:DatatypeProperty ;
        rdfs:range xsd:boolean ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "True if the image is interlaced, false if not." ;
        rdfs:label "interlaceMode" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#isPasswordProtected
nfo:isPasswordProtected a owl:DatatypeProperty ;
        rdfs:domain nfo:ArchiveItem ;
        rdfs:range xsd:boolean ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "States if a given resource is password-protected." ;
        rdfs:label "isPasswordProtected" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#lfeChannels
nfo:lfeChannels a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:channels ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Number of Low Frequency Expansion (subwoofer) channels." ;
        rdfs:label "lfeChannels" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#lineCount
nfo:lineCount a owl:DatatypeProperty ;
        rdfs:domain nfo:TextDocument ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of lines in a text document" ;
        rdfs:label "lineCount" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#occupiedSpace
nfo:occupiedSpace a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Occupied storage space of the filesystem." ;
        rdfs:label "occupiedSpace" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#originalLocation
nfo:originalLocation a owl:DatatypeProperty ;
        rdfs:domain nfo:DeletedResource ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The original location of the deleted resource." ;
        rdfs:label "originalLocation" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#pageCount
nfo:pageCount a owl:DatatypeProperty ;
        rdfs:domain nfo:PaginatedTextDocument ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Number of pages." ;
        rdfs:label "pageCount" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#pageNumber
nfo:pageNumber a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Page linked by the bookmark." ;
        rdfs:label "pageNumber" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#paletteSize
nfo:paletteSize a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The number of colors defined in palette of the raster image." ;
        rdfs:label "palette size" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#permissions
nfo:permissions a owl:DatatypeProperty ;
        rdfs:domain nfo:FileDataObject ;
        rdfs:range xsd:string ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "A string containing the permissions of a file. A feature common in many UNIX-like
operating systems." ;
        rdfs:label "permissions" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#programmingLanguage
nfo:programmingLanguage a owl:DatatypeProperty ;
        rdfs:domain nfo:SourceCode ;
        rdfs:range xsd:string ;
        rdfs:comment "Indicates the name of the programming language this source code file is written in.
Examples might include 'C', 'C++', 'Java' etc." ;
        rdfs:label "programmingLanguage" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#rate
nfo:rate a owl:DatatypeProperty ;
        rdfs:range xsd:float ;
        rdfs:comment "A common superproperty for all properties specifying the media rate. Examples of
subproperties may include frameRate for video and sampleRate for audio. This property is expressed in units
per second." ;
        rdfs:label "rate" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#rearChannels
nfo:rearChannels a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:channels ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Number of rear channels." ;
        rdfs:label "rearChannels" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#sampleCount
nfo:sampleCount a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:count ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of samples in an audio clip." ;
        rdfs:label "sampleCount" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#sampleRate
nfo:sampleRate a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:rate ;
        rdfs:range xsd:float ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of audio samples per second." ;
        rdfs:label "sampleRate" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#sideChannels
nfo:sideChannels a owl:DatatypeProperty ;
        rdfs:subPropertyOf nfo:channels ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Number of side channels" ;
        rdfs:label "sideChannels" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#streamPosition
nfo:streamPosition a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Stream position of the bookmark, suitable for e.g. audio books. Expressed in
milliseconds" ;
        rdfs:label "streamPosition" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#totalSpace
nfo:totalSpace a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Total storage space of the filesystem, which can be different from nie:contentSize
because the latter includes filesystem format overhead." ;
        rdfs:label "totalSpace" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#uncompressedSize
nfo:uncompressedSize a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Uncompressed size of the content of a compressed file." ;
        rdfs:label "uncompressedSize" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#uuid
nfo:uuid a owl:DatatypeProperty .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#verticalResolution
nfo:verticalResolution a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Vertical resolution of an Image (if printed). Expressed in DPI" ;
        rdfs:label "verticalResolution" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#width
nfo:width a owl:DatatypeProperty ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "Visual content width in pixels." ;
        rdfs:label "width" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#wordCount
nfo:wordCount a owl:DatatypeProperty ;
        rdfs:domain nfo:TextDocument ;
        rdfs:range xsd:integer ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The amount of words in a text document." ;
        rdfs:label "wordCount" .
# http://www.w3.org/ns/prov#endedAtTime
prov1:endedAtTime a owl:DatatypeProperty .
# http://www.w3.org/ns/prov#startedAtTime
prov1:startedAtTime a owl:DatatypeProperty .
# #    Classes
# http://purl.org/dc/terms/Policy
terms:Policy a owl:Class ;
        rdfs:subClassOf :Proposition .
# http://purl.org/net/cosi#AcademicDiscipline
:AcademicDiscipline a owl:Class ;
        rdfs:subClassOf :Proposition ;
        rdfs:comment "SUMO \"An academic or applied discipline with recognized experts and with a core of
accepted theory or practice. Note that FieldOfStudy is a subclass of Proposition, because a FieldOfStudy is
understood to be a body of abstract, informational content, with varying degrees of certainty attached to
each element of this content.\"" .
```

```
# http://purl.org/net/cosi#Article
:Article a owl:Class ;
        rdfs:subClassOf :ScholarCommunication ;
        prov1:definition "A relatively short Text that either is unbound or is bound with other Articles in
a Book"^^xsd:string .
# http://purl.org/net/cosi#Book
:Book a owl:Class ;
        rdfs:subClassOf :ScholarCommunication .
# http://purl.org/net/cosi#Computer
:Computer a owl:Class ;
        rdfs:subClassOf :DigitalInstrument .
# http://purl.org/net/cosi#Conclusion
:Conclusion a owl:Class ;
        rdfs:subClassOf :Proposition .
# http://purl.org/net/cosi#CopyRightInformation
:CopyRightInformation a owl:Class ;
        rdfs:subClassOf :RightsStatement ;
        rdfs:comment "When rights basis is a copyright, copyrightInformation should be
provided."^^xsd:string ;
        prov1:definition "Information about the copyright status of the object(s)."^^xsd:string .
# http://purl.org/net/cosi#DigitalInstrument
:DigitalInstrument a owl:Class ;
        rdfs:subClassOf :Instrument ;
        rdfs:comment "A Digital Instrument is a device based on a digital computational system."@en ;
        rdfs:label "Digital Instrument"@en .
# http://purl.org/net/cosi#ErrorInComparison
:ErrorInComparison a owl:Class ;
        rdfs:subClassOf :ResultError .
# http://purl.org/net/cosi#ErrorInConclusion
:ErrorInConclusion a owl:Class ;
        rdfs:subClassOf :ResultError .
# http://purl.org/net/cosi#ErrorIncompleteData
:ErrorIncompleteData a owl:Class ;
        rdfs:subClassOf :ResultError .
# http://purl.org/net/cosi#Experiment
:Experiment a owl:Class ;
        rdfs:subClassOf :Investigation ;
        rdfs:comment "Experiments. Investigations into the physical behaviour of the environment usually to
test a hypothesis, typically involving an instrument operating under some instrumental settings and
environmental conditions, and generating datasets in files."@en ;
        rdfs:label "Experiment"@en ;
        prov1:definition "The experiment relies in a concrete controlled environment, usually a laboratory
or an experiment site"@en .
# http://purl.org/net/cosi#ExperimentProcedure
:ExperimentProcedure a owl:Class ;
        rdfs:subClassOf :InvestigationProcedure ;
        rdfs:comment "An Experiment Procedure is a procedure or a set of ordered procedures involving
artefacts of an experiment and their interoperable functionality"@en ;
        rdfs:label "Experiment Procedure"@en ;
        prov1:definition "An Experiment Procedure is a procedure or a set of ordered procedures involving
artefacts of an experiment and their interoperable functionality"@en .
# http://purl.org/net/cosi#Hypothesis
:Hypothesis a owl:Class ;
        rdfs:subClassOf :Proposition .
# http://purl.org/net/cosi#HypothesisAcceptance
:HypothesisAcceptance a owl:Class ;
        rdfs:subClassOf :HypothesisStatement .
# http://purl.org/net/cosi#HypothesisRejection
:HypothesisRejection a owl:Class ;
        rdfs:subClassOf :HypothesisStatement .
# http://purl.org/net/cosi#HypothesisStatement
:HypothesisStatement a owl:Class ;
        rdfs:subClassOf :Proposition .
# http://purl.org/net/cosi#Input
:Input a owl:Class ;
        rdfs:subClassOf nie:DataObject ;
        dc:description "An Input is what is put in, taken in, or operated on by any process or system"@en ;
        rdfs:label "Input"@en ;
        prov1:definition "A dataobject operated on by any process or activity"@en .
# http://purl.org/net/cosi#Institution
:Institution a owl:Class ;
        owl:equivalentClass prov1:Organization ;
        rdfs:comment "An administrative organization, same as prov:Organization"@en ;
        rdfs:label "Institution"@en .
# http://purl.org/net/cosi#Instrument
:Instrument a owl:Class ;
        rdfs:comment "An Instrument is a device used in investigations"@en ;
        rdfs:label "Instrument"@en ;
        prov1:definition "An Instrument is a device used in investigations"@en .
# http://purl.org/net/cosi#Investigation
:Investigation a owl:Class ;
        owl:equivalentClass :S5_Investigation ;
        rdfs:subClassOf :Process , _:genid1 .
_:genid1 a owl:Restriction ;
        owl:onProperty :hasInvestigator ;
        owl:someValuesFrom :Investigator .
:Investigation rdfs:subClassOf _:genid2 .
_:genid2 a owl:Restriction ;
```

```
        owl:onProperty :hasProcedure ;
        owl:someValuesFrom :Procedure .
:Investigation rdfs:subClassOf _:genid3 .
_:genid3 a owl:Restriction ;
        owl:onProperty :hasInvestigator ;
        owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
        owl:onClass :PrincipalInvestigator .
:Investigation rdfs:subClassOf _:genid4 .
_:genid4 a owl:Restriction ;
        owl:onProperty :hasHypothesis ;
        owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger ;
        owl:onClass :Hypothesis .
:Investigation rdfs:label "Investigation"@en ;
        prov1:definition """Definition 2: Investigation is an examination process in support of a
hypothesis. It is part of a systematic study and adheres to a scientific procedure
An investigation has a title, a description and is identifiable. It has at least a principal investigator
role and an undefined number of additional investigators"""@en .
# http://purl.org/net/cosi#InvestigationProcedure
:InvestigationProcedure a owl:Class ;
        rdfs:subClassOf :Procedure ;
        rdfs:comment "An Investigation Procedure is a well defined Procedure related to an Investigation
Process"@en ;
        rdfs:label "Investigation Procedure"@en .
# http://purl.org/net/cosi#Investigator
:Investigator a owl:Class ;
        rdfs:subClassOf prov1:Role ;
        dc:description "An individual who is involved in a research process"@en ;
        rdfs:label "Investigator"@en ;
        prov1:definition "An Investigator is an entity involved in a research process"@en .
# http://purl.org/net/cosi#Journal
:Journal a owl:Class ;
        rdfs:subClassOf :ScholarCommunication .
# http://purl.org/net/cosi#LicenceInformation
:LicenceInformation a owl:Class ;
        rdfs:subClassOf :RightsStatement ;
        rdfs:comment "When rights basis is a license, licenseInformation should be provided."^^xsd:string ;
        prov1:definition "Information about a license or other agreement granting permissions related to an
object."^^xsd:string .
# http://purl.org/net/cosi#Magazine
:Magazine a owl:Class ;
        rdfs:subClassOf :ScholarCommunication .
# http://purl.org/net/cosi#Measurement
:Measurement a owl:Class ;
        rdfs:subClassOf :Investigation ;
        rdfs:comment "Measurements. Investigations that record the state of some aspect of the environment
over a sequence of points in time and space, using some passive detector, e.g., the measurement of
temperature at a point on the earth surface taken hourly using a thermometer of known accuracy."@en ;
        rdfs:label "Measurement"@en ;
        prov1:definition "A Measurement is a form of investigation defined by a Measurement Procedure"@en .
# http://purl.org/net/cosi#Motivation_Publication
:Motivation_Publication a owl:Class ;
        rdfs:subClassOf :Publication .
# http://purl.org/net/cosi#Observation
:Observation a owl:Class ;
        rdfs:subClassOf :Investigation , _:genid5 .
_:genid5 a owl:Restriction ;
        owl:onProperty :hasObservationProcedure ;
        owl:minQualifiedCardinality "1"^^xsd:nonNegativeInteger ;
        owl:onClass :ObservationProcedure .
:Observation dc:description "An Observation is a form of investigation defined by an Observation Procedure
and run by one or more observers."@en ;
        rdfs:label "Observation"@en ;
        prov1:definition "An Investigation Activity where facts are learned by a (human) observer."@en .
# http://purl.org/net/cosi#ObservationProcedure
:ObservationProcedure a owl:Class ;
        rdfs:subClassOf :InvestigationProcedure ;
        rdfs:label "Observation Procedure"@en ;
        prov1:definition "An Observation Procedure is run by one or more observers and the gathered
information is result of human perception"@en .
# http://purl.org/net/cosi#Observer
:Observer a owl:Class ;
        rdfs:subClassOf :Investigator , prov1:Person , _:genid6 .
_:genid6 a owl:Restriction ;
        owl:onProperty :isObserverOf ;
        owl:someValuesFrom :ObservationProcedure .
:Observer dc:description "A person who becomes aware of through the senses"@en ;
        rdfs:label "Observer"@en ;
        prov1:definition "An Observer is a person involved in some Observation process who becomes aware of
through the senses."@en .
# http://purl.org/net/cosi#Onymous
:Onymous a owl:Class ;
        rdfs:subClassOf _:genid7 .
_:genid7 a owl:Restriction ;
        owl:onProperty :hasIdentifier ;
        owl:cardinality "1"^^xsd:nonNegativeInteger .
:Onymous rdfs:subClassOf _:genid8 .
_:genid8 a owl:Restriction ;
        owl:onProperty :hasIdentifierValue ;
        owl:cardinality "1"^^xsd:nonNegativeInteger .
:Onymous dc:description "An Onymous is an entity that is identifiable, has a title and a description"@en ;
        rdfs:comment "An Onymous is an entity that is identifiable, has a title and a description"@en .
# http://purl.org/net/cosi#Output
:Output a owl:Class ;
        rdfs:subClassOf nie:DataObject ;
        dc:description "An output is the result of a process."@en ;
        rdfs:label "Output"@en ;
        prov1:definition "An output is the dataobject resuling of a process or activity"@en .
# http://purl.org/net/cosi#Parameter
:Parameter a owl:Class ;
        rdfs:subClassOf nie:DataObject .
# http://purl.org/net/cosi#PersistendIdentifier
:PersistendIdentifier a owl:Class ;
        rdfs:subClassOf :Onymous , _:genid9 .
_:genid9 a owl:Restriction ;
        owl:onProperty :hasHandler ;
        owl:cardinality "1"^^xsd:nonNegativeInteger .
# http://purl.org/net/cosi#Preservation
:Preservation a owl:Class ;
        rdfs:subClassOf prov1:Activity ;
        prov1:definition "A preservation action is the definition of sets of neccessary information created
upon ingest of a digital object (eg Result Set from an Investigation) in a repository."^^xsd:string .
# http://purl.org/net/cosi#PreservationPlan
:PreservationPlan a owl:Class ;
        rdfs:subClassOf :Preservation ;
        prov1:definition "The systematic process that defines the goals and priorities for the
preservation"^^xsd:string .
# http://purl.org/net/cosi#PrincipalInvestigator
:PrincipalInvestigator a owl:Class ;
        rdfs:subClassOf :Investigator ;
        dc:description "An investigator responsible for a specific Investigation"@en , "Principal
Investigator"@en ;
        prov1:definition "An investigator which is the responsible for a specific Investigation"@en .
# http://purl.org/net/cosi#Procedure
:Procedure a owl:Class ;
        rdfs:label "Procedure"@en ;
        prov1:definition """Procedure is a specific was to carry out an activity or a process
An Investigation Procedure \"an accepted or approved\" procedure in the domain of an
investigation"""^^rdfs:Literal .
# http://purl.org/net/cosi#Proceedings
:Proceedings a owl:Class ;
        rdfs:subClassOf :ScholarCommunication .
# http://purl.org/net/cosi#Process
:Process a owl:Class ;
        owl:equivalentClass prov1:Activity ;
        dc:description "A process is an activity or a set of activities that use resources to transfer
input to output"@en ;
        rdfs:label "Process"@en ;
        prov1:definition "A process is an activity or a set of activities that use resources to transfer
input to output"@en .
# http://purl.org/net/cosi#Programme
:Programme a owl:Class ;
        rdfs:subClassOf :Proposition ;
        rdfs:comment """Scientific research programme
Programmes related studies that have a common theme which are usually funded and resourced directly or with
an intermediary organisation under the rubric of the programme."""@en ;
        rdfs:label "Programme"@en .
# http://purl.org/net/cosi#Proposition
:Proposition a owl:Class ;
        prov1:definition "Propositions are Abstract entities that express a complete thought or a set of
such thoughts."^^xsd:string .
# http://purl.org/net/cosi#Publication
:Publication a owl:Class ;
        owl:equivalentClass :ScholarCommunication ;
        rdfs:subClassOf nfo:Document ;
        rdfs:comment "publication (the communication of something to the public; making information
generally known)"@en .
# http://purl.org/net/cosi#ResearchFacility
:ResearchFacility a owl:Class ;
        dc:description "An entity where a particular process can be run. In the context of this ontology,
the Research Facility is used to denote a geo-located artifact  such as a Laboratory or research
facility."@en ;
        rdfs:label "Research Facility"@en ;
        prov1:definition "A stationary artifact where particular process can be run"@en .
# http://purl.org/net/cosi#Result
:Result a owl:Class ;
        nao:prefLabel "Result Set"@en ;
        rdfs:comment "Results is the representation of an activity once it has ended"@en ;
        rdfs:label "Result"@en ;
        prov1:definition "A result is collection of data objects gathered in the execution of an activity.
It consists of a data collection and a descriptive document on the findings"@en .
# http://purl.org/net/cosi#ResultError
:ResultError a owl:Class ;
        rdfs:subClassOf :Proposition .
# http://purl.org/net/cosi#ResultSet
:ResultSet a owl:Class ;
        rdfs:subClassOf :Result , nie:DataObject ;
        prov1:definition "The set of data generated as a result of an investigation operation"^^xsd:string
.
# http://purl.org/net/cosi#Result_Publication
:Result_Publication a owl:Class ;
        rdfs:subClassOf :Publication .
# http://purl.org/net/cosi#Rig
:Rig a owl:Class ;
        owl:equivalentClass core:OrderedList ;
        rdfs:comment "A RIG is an ordered combination of all the artefacts participating in an
investigation"@en ;
        rdfs:label "Rig"@en ;
        prov1:definition "A RIG is an ordered combination of all the artefacts participating in an
investigation"@en .
# http://purl.org/net/cosi#RightsDocumentation
:RightsDocumentation a owl:Class ;
        rdfs:subClassOf nie:DataObject , nie:InformationElement ;
        prov1:definition "A designation used to uniquely identify documentation supporting the specified
rights within the repository system."^^xsd:string .
# http://purl.org/net/cosi#RightsStatement
:RightsStatement a owl:Class ;
        rdfs:subClassOf nie:InformationElement ;
        rdfs:comment "This semantic unit is optional because in some cases rights may be unknown.
Institutions are encouraged to record rights information when possible. Either rightsStatement or
rightsExtension must be present if the Rights entity is included. The rightsStatement should be repeated
when the act(s) described has more than one basis, or when different acts have different
bases."^^xsd:string ;
        prov1:definition "Documentation of the repository's right to perform one or more acts."^^xsd:string
.
# http://purl.org/net/cosi#S1_ScientificTaskDefinition
:S1_ScientificTaskDefinition a owl:Class ;
        rdfs:subClassOf :ScientificTask ;
        rdfs:comment """Define the question, Define the idea
Search the literature, formulate statement""" .
# http://purl.org/net/cosi#S2_RelatedInformation
:S2_RelatedInformation a owl:Class ;
        rdfs:subClassOf :ScientificTask ;
        prov1:definition "Related Information points to information gathering is the scientific task of
collecting information for further analysis or for other scientific purposes."@en .
# http://purl.org/net/cosi#S3_HypothesisDefinition
:S3_HypothesisDefinition a owl:Class ;
        rdfs:subClassOf :ScientificTask .
# http://purl.org/net/cosi#S4_MethodologyDefinition
:S4_MethodologyDefinition a owl:Class ;
        rdfs:subClassOf :ScientificTask ;
        prov1:definition "Definition of a methodology and procedure to be followed." .
# http://purl.org/net/cosi#S5_Investigation
:S5_Investigation a owl:Class ;
        rdfs:subClassOf :ScientificTask ;
        prov1:definition "Execution of an Investigation or subclass of an Investigation" .
# http://purl.org/net/cosi#S6_Conclusion
:S6_Conclusion a owl:Class ;
        rdfs:subClassOf :ScientificTask ;
        prov1:definition "Interpretation of the Results and challenge of the Hypothesis." .
# http://purl.org/net/cosi#S7_ResultPublication
:S7_ResultPublication a owl:Class ;
        rdfs:subClassOf :ScientificTask ;
        prov1:definition "Reference to the publication of the results. This may be merely in a repository
or in a classical scholarly communication format" .
# http://purl.org/net/cosi#ScholarCommunication
:ScholarCommunication a owl:Class ;
        rdfs:subClassOf nie:InformationElement , nfo:Document .
# http://purl.org/net/cosi#ScientificTask
:ScientificTask a owl:Class ;
        rdfs:subClassOf :Proposition ;
        rdfs:comment "A scientific task follows a scientific method in a study"^^xsd:string .
# http://purl.org/net/cosi#Simulation
:Simulation a owl:Class ;
        rdfs:subClassOf :Investigation ;
        rdfs:comment "A Simulation is a form of investigation created on a computer imitation of a real
scenario based on digital model"@en ;
        rdfs:label "Simulation"@en ;
        prov1:definition "A Simulation is a form of investigation created on a computer imitation of a real
scenario based on digital model"@en .
# http://purl.org/net/cosi#Study
:Study a owl:Class ;
        rdfs:subClassOf :Process ;
        rdfs:comment "Studies investigate some aspect of science"@en ;
        rdfs:label "Study"@en .
# http://purl.org/net/cosi#Summary
:Summary a owl:Class ;
        rdfs:subClassOf :ScholarCommunication .
# http://purl.org/net/cosi#TechnicalPaper
:TechnicalPaper a owl:Class ;
        rdfs:subClassOf :ScholarCommunication .
# http://purl.org/net/cosi#Topic
:Topic a owl:Class ;
        rdfs:subClassOf :Proposition ;
        rdfs:comment "The topic object property provides a subject relevant to the particular study.
Considering the focus of the model to be more investigation and result set oriented, the topic entity might
be designed minimalistic as defined in the CCLRC Scientific Metadata Model. Such a Topic will include a set
of keywords and a subject pointing to specific disciplines (yet another another entity). However, it makes
more sense to connect the \"topic\" with a standard vocabulary which contains a complete list of scientific
disciplines. Such a vocabulary might the German standard classification described in the “DFG
Classification of Subject Area, Research Area and Scientific Discipline” [36]. A SKOS representation of
this vocabulary has been also defined for the context of this research and is presented in URL-TODOFILL"@en
;
        rdfs:label "Topic"@en .
# http://purl.org/ontology/olo/core#OrderedList
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#DataObject
nie:DataObject a owl:Class .
# http://www.semanticdesktop.org/ontologies/2007/01/19/nie#InformationElement
nie:InformationElement a owl:Class .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#ArchiveItem
nfo:ArchiveItem a owl:Class ;
        rdfs:subClassOf nfo:EmbeddedFileDataObject ;
        rdfs:comment "A file entity inside an archive." ;
        rdfs:label "ArchiveItem" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Attachment
nfo:Attachment a owl:Class ;
        rdfs:subClassOf nfo:EmbeddedFileDataObject ;
        rdfs:comment "A file attached to another data object. Many data formats allow for attachments:
emails, vcards, ical events, id3 and exif..." ;
        rdfs:label "Attachment" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#DataContainer
nfo:DataContainer a owl:Class ;
        rdfs:subClassOf nie:InformationElement ;
        rdfs:comment "A superclass for all entities, whose primary purpose is to serve as containers for
other data object. They usually don't have any \"meaning\" by themselves. Examples include folders,
archives and optical disc images." ;
        rdfs:label "DataContainer" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#DeletedResource
nfo:DeletedResource a owl:Class ;
        rdfs:subClassOf nfo:FileDataObject ;
        rdfs:comment "A file entity that has been deleted from the original source. Usually such entities
are stored within various kinds of 'Trash' or 'Recycle Bin' folders." ;
        rdfs:label "DeletedResource" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Document
nfo:Document a owl:Class ;
        rdfs:subClassOf nie:InformationElement ;
        rdfs:comment "A generic document. A common superclass for all documents on the desktop." ;
        rdfs:label "Document" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#EmbeddedFileDataObject
nfo:EmbeddedFileDataObject a owl:Class ;
        rdfs:subClassOf nfo:FileDataObject ;
        rdfs:comment "A file embedded in another data object. There are many ways in which a file may be
embedded in another one. Use this class directly only in cases if none of the subclasses gives a better
description of your case." ;
        rdfs:label "EmbeddedFileDataObject" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#FileDataObject
nfo:FileDataObject a owl:Class ;
        rdfs:subClassOf nie:DataObject ;
        rdfs:comment "A resource containing a finite sequence of bytes with arbitrary information, that is
available to a computer program and is usually based on some kind of durable storage. A file is durable in
the sense that it remains available for programs to use after the current program has finished." ;
        rdfs:label "file" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Folder
nfo:Folder a owl:Class ;
        rdfs:subClassOf nfo:DataContainer ;
        rdfs:comment "A folder/directory. Examples of folders include folders on a filesystem and message
folders in a mailbox." ;
        rdfs:label "Folder" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#HtmlDocument
nfo:HtmlDocument a owl:Class ;
        rdfs:subClassOf nfo:PlainTextDocument ;
        rdfs:comment "A HTML document, may contain links to other files." ;
        rdfs:label "HtmlDocument" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#LocalFileDataObject
nfo:LocalFileDataObject a owl:Class ;
        rdfs:subClassOf nfo:FileDataObject ;
        rdfs:comment "A local file data object which is stored on a local file system. Its nie:url always
uses the file:/ protocol. The main use of this class is to distinguish local and non-local files." ;
        rdfs:label "local file" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#PaginatedTextDocument
nfo:PaginatedTextDocument a owl:Class ;
        rdfs:subClassOf nfo:TextDocument ;
        rdfs:comment "A file containing a text document, that is unambiguously divided into pages. Examples
might include PDF, DOC, PS, DVI etc." ;
        rdfs:label "PaginatedTextDocument" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#PlainTextDocument
nfo:PlainTextDocument a owl:Class ;
        rdfs:subClassOf nfo:TextDocument ;
        rdfs:comment "A file containing plain text (ASCII, Unicode or other encodings). Examples may
include TXT, HTML, XML, program source code etc." ;
        rdfs:label "PlainTextDocument" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Presentation
```

```
nfo:Presentation a owl:Class ;
        rdfs:subClassOf nfo:Document ;
        rdfs:comment "A Presentation made by some presentation software (Corel Presentations, OpenOffice
Impress, MS Powerpoint etc.)" ;
        rdfs:label "Presentation" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#RemoteDataObject
nfo:RemoteDataObject a owl:Class ;
        rdfs:subClassOf nfo:FileDataObject ;
        rdfs:comment "A file data object stored at a remote location. Don't confuse this class with a
RemotePortAddress. This one applies to a particular resource, RemotePortAddress applies to an address, that
can have various interpretations." ;
        rdfs:label "RemoteDataObject" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#SourceCode
nfo:SourceCode a owl:Class ;
        rdfs:subClassOf nfo:PlainTextDocument ;
        rdfs:comment "Code in a compilable or interpreted programming language." ;
        rdfs:label "SourceCode" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Spreadsheet
nfo:Spreadsheet a owl:Class ;
        rdfs:subClassOf nfo:Document ;
        rdfs:comment "A spreadsheet, created by a spreadsheet application. Examples might include Gnumeric,
OpenOffice Calc or MS Excel." ;
        rdfs:label "Spreadsheet" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#TextDocument
nfo:TextDocument a owl:Class ;
        rdfs:subClassOf nfo:Document ;
        rdfs:comment "A text document" ;
        rdfs:label "TextDocument" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Website
nfo:Website a owl:Class ;
        rdfs:subClassOf nie:InformationElement ;
        rdfs:comment "A website, usually a container for remote resources, that may be interpreted as
HTMLDocuments, images or other types of content." ;
        rdfs:label "Website" .
# http://www.w3.org/2000/01/rdf-schema#Class
rdfs:Class a owl:Class .
# http://www.w3.org/2002/07/owl#Thing
owl:Thing a owl:Class .
# http://www.w3.org/ns/prov#Activity
prov1:Activity a owl:Class .
# http://www.w3.org/ns/prov#Entity
prov1:Entity a owl:Class .
# http://www.w3.org/ns/prov#Organization
# http://www.w3.org/ns/prov#Role
prov1:Role a owl:Class .
# #    Individuals
# http://purl.org/dc/aboutdcmi#DCMI
<http://purl.org/dc/aboutdcmi#DCMI> a owl:NamedIndividual .
# http://purl.org/dc/terms/
<http://purl.org/dc/terms/> a owl:NamedIndividual ;
        terms:modified "2012-06-14"^^xsd:date ;
        terms:publisher <http://purl.org/dc/aboutdcmi#DCMI> ;
        terms:title "DCMI Metadata Terms - other"@en .
# http://purl.org/dc/terms/Agent
terms:Agent a owl:NamedIndividual .
# http://purl.org/dc/terms/AgentClass
terms:AgentClass a owl:NamedIndividual .
# http://purl.org/dc/terms/BibliographicResource
terms:BibliographicResource a owl:NamedIndividual .
# http://purl.org/dc/terms/Box
terms:Box a owl:NamedIndividual .
# http://purl.org/dc/terms/DCMIType
terms:DCMIType a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#DCMIType-005> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2012-06-14"^^xsd:date ;
        rdfs:comment "The set of classes specified by the DCMI Type Vocabulary, used to categorize the
nature or genre of the resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "DCMI Type Vocabulary"@en ;
        rdfs:seeAlso <http://purl.org/dc/dcmitype/> .
# http://purl.org/dc/terms/DDC
terms:DDC a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#DDC-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of conceptual resources specified by the Dewey Decimal Classification."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "DDC"@en ;
        rdfs:seeAlso <http://www.oclc.org/dewey/> .
# http://purl.org/dc/terms/FileFormat
terms:FileFormat a owl:NamedIndividual .
# http://purl.org/dc/terms/Frequency
terms:Frequency a owl:NamedIndividual .
# http://purl.org/dc/terms/IMT
terms:IMT a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#IMT-004> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of media types specified by the Internet Assigned Numbers Authority."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "IMT"@en ;
        rdfs:seeAlso <http://www.iana.org/assignments/media-types/> .
# http://purl.org/dc/terms/ISO3166
terms:ISO3166 a owl:NamedIndividual .
# http://purl.org/dc/terms/ISO639-2
terms:ISO639-2 a owl:NamedIndividual .
# http://purl.org/dc/terms/ISO639-3
terms:ISO639-3 a owl:NamedIndividual .
# http://purl.org/dc/terms/Jurisdiction
terms:Jurisdiction a owl:NamedIndividual .
# http://purl.org/dc/terms/LCC
terms:LCC a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#LCC-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of conceptual resources specified by the Library of Congress
Classification."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "LCC"@en ;
        rdfs:seeAlso <http://lcweb.loc.gov/catdir/cpso/lcco/lcco.html> .
# http://purl.org/dc/terms/LCSH
terms:LCSH a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#LCSH-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of labeled concepts specified by the Library of Congress Subject
Headings."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "LCSH"@en .
# http://purl.org/dc/terms/LicenseDocument
terms:LicenseDocument a owl:NamedIndividual .
# http://purl.org/dc/terms/LinguisticSystem
terms:LinguisticSystem a owl:NamedIndividual .
# http://purl.org/dc/terms/Location
terms:Location a owl:NamedIndividual .
# http://purl.org/dc/terms/LocationPeriodOrJurisdiction
terms:LocationPeriodOrJurisdiction a owl:NamedIndividual .
# http://purl.org/dc/terms/MESH
terms:MESH a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#MESH-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of labeled concepts specified by the Medical Subject Headings."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "MeSH"@en ;
        rdfs:seeAlso <http://www.nlm.nih.gov/mesh/meshhome.html> .
# http://purl.org/dc/terms/MediaType
terms:MediaType a owl:NamedIndividual .
# http://purl.org/dc/terms/MediaTypeOrExtent
terms:MediaTypeOrExtent a owl:NamedIndividual .
# http://purl.org/dc/terms/MethodOfAccrual
terms:MethodOfAccrual a owl:NamedIndividual .
# http://purl.org/dc/terms/MethodOfInstruction
terms:MethodOfInstruction a owl:NamedIndividual .
# http://purl.org/dc/terms/NLM
terms:NLM a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#NLM-002> ;
        terms:issued "2005-06-13"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of conceptual resources specified by the National Library of Medicine
Classification."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "NLM"@en ;
        rdfs:seeAlso <http://wwwcf.nlm.nih.gov/class/> .
# http://purl.org/dc/terms/Period
terms:Period a owl:NamedIndividual .
# http://purl.org/dc/terms/PeriodOfTime
terms:PeriodOfTime a owl:NamedIndividual .
# http://purl.org/dc/terms/PhysicalMedium
terms:PhysicalMedium a owl:NamedIndividual .
# http://purl.org/dc/terms/PhysicalResource
terms:PhysicalResource a owl:NamedIndividual .
# http://purl.org/dc/terms/Point
terms:Point a owl:NamedIndividual .
# http://purl.org/dc/terms/Policy
terms:Policy a owl:NamedIndividual .
# http://purl.org/dc/terms/ProvenanceStatement
terms:ProvenanceStatement a owl:NamedIndividual .
# http://purl.org/dc/terms/RFC1766
terms:RFC1766 a owl:NamedIndividual .
# http://purl.org/dc/terms/RFC3066
terms:RFC3066 a owl:NamedIndividual .
# http://purl.org/dc/terms/RFC4646
terms:RFC4646 a owl:NamedIndividual .
# http://purl.org/dc/terms/RFC5646
```

```
terms:RFC5646 a owl:NamedIndividual .
# http://purl.org/dc/terms/RightsStatement
terms:RightsStatement a owl:NamedIndividual ;
        rdfs:comment "This semantic unit is optional because in some cases rights may be unknown.
Institutions are encouraged to record rights information when possible."^^xsd:string ;
        prov1:definition "Documentation of the repository's right to perform one or more acts."@en .
# http://purl.org/dc/terms/SizeOrDuration
terms:SizeOrDuration a owl:NamedIndividual .
# http://purl.org/dc/terms/Standard
terms:Standard a owl:NamedIndividual .
# http://purl.org/dc/terms/TGN
terms:TGN a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#TGN-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of places specified by the Getty Thesaurus of Geographic Names."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "TGN"@en ;
        rdfs:seeAlso <http://www.getty.edu/research/tools/vocabulary/tgn/index.html> .
# http://purl.org/dc/terms/UDC
terms:UDC a owl:NamedIndividual ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#UDC-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The set of conceptual resources specified by the Universal Decimal
Classification."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "UDC"@en ;
        rdfs:seeAlso <http://www.udcc.org/> .
# http://purl.org/dc/terms/URI
terms:URI a owl:NamedIndividual .
# http://purl.org/dc/terms/W3CDTF
terms:W3CDTF a owl:NamedIndividual .
# http://purl.org/dc/terms/abstract
terms:abstract a owl:NamedIndividual .
# http://purl.org/dc/terms/accessRights
terms:accessRights a owl:NamedIndividual .
# http://purl.org/dc/terms/accrualMethod
terms:accrualMethod a owl:NamedIndividual .
# http://purl.org/dc/terms/accrualPeriodicity
terms:accrualPeriodicity a owl:NamedIndividual .
# http://purl.org/dc/terms/accrualPolicy
terms:accrualPolicy a owl:NamedIndividual .
# http://purl.org/dc/terms/alternative
terms:alternative a owl:NamedIndividual .
# http://purl.org/dc/terms/audience
terms:audience a owl:NamedIndividual .
# http://purl.org/dc/terms/available
terms:available a owl:NamedIndividual .
# http://purl.org/dc/terms/bibliographicCitation
terms:bibliographicCitation a owl:NamedIndividual .
# http://purl.org/dc/terms/conformsTo
terms:conformsTo a owl:NamedIndividual .
# http://purl.org/dc/terms/contributor
terms:contributor a owl:NamedIndividual .
# http://purl.org/dc/terms/coverage
terms:coverage a owl:NamedIndividual .
# http://purl.org/dc/terms/created
terms:created a owl:NamedIndividual .
# http://purl.org/dc/terms/creator
terms:creator a owl:NamedIndividual .
# http://purl.org/dc/terms/date
terms:date a owl:NamedIndividual .
# http://purl.org/dc/terms/dateAccepted
terms:dateAccepted a owl:NamedIndividual .
# http://purl.org/dc/terms/dateCopyrighted
terms:dateCopyrighted a owl:NamedIndividual .
# http://purl.org/dc/terms/dateSubmitted
terms:dateSubmitted a owl:NamedIndividual .
# http://purl.org/dc/terms/description
terms:description a owl:NamedIndividual .
# http://purl.org/dc/terms/educationLevel
terms:educationLevel a owl:NamedIndividual .
# http://purl.org/dc/terms/extent
terms:extent a owl:NamedIndividual .
# http://purl.org/dc/terms/format
terms:format a owl:NamedIndividual .
# http://purl.org/dc/terms/hasFormat
terms:hasFormat a owl:NamedIndividual .
# http://purl.org/dc/terms/hasPart
terms:hasPart a owl:NamedIndividual .
# http://purl.org/dc/terms/hasVersion
terms:hasVersion a owl:NamedIndividual .
# http://purl.org/dc/terms/identifier
terms:identifier a owl:NamedIndividual .
# http://purl.org/dc/terms/instructionalMethod
terms:instructionalMethod a owl:NamedIndividual .
# http://purl.org/dc/terms/isFormatOf
```

```
terms:isFormatOf a owl:NamedIndividual .
# http://purl.org/dc/terms/isPartOf
terms:isPartOf a owl:NamedIndividual .
# http://purl.org/dc/terms/isReferencedBy
terms:isReferencedBy a owl:NamedIndividual .
# http://purl.org/dc/terms/isReplacedBy
terms:isReplacedBy a owl:NamedIndividual .
# http://purl.org/dc/terms/isRequiredBy
terms:isRequiredBy a owl:NamedIndividual .
# http://purl.org/dc/terms/isVersionOf
terms:isVersionOf a owl:NamedIndividual .
# http://purl.org/dc/terms/issued
terms:issued a owl:NamedIndividual .
# http://purl.org/dc/terms/language
terms:language a owl:NamedIndividual .
# http://purl.org/dc/terms/license
terms:license a owl:NamedIndividual .
# http://purl.org/dc/terms/mediator
terms:mediator a owl:NamedIndividual .
# http://purl.org/dc/terms/medium
terms:medium a owl:NamedIndividual .
# http://purl.org/dc/terms/modified
terms:modified a owl:NamedIndividual .
# http://purl.org/dc/terms/provenance
terms:provenance a owl:NamedIndividual .
# http://purl.org/dc/terms/publisher
terms:publisher a owl:NamedIndividual .
# http://purl.org/dc/terms/references
terms:references a owl:NamedIndividual .
# http://purl.org/dc/terms/relation
terms:relation a owl:NamedIndividual .
# http://purl.org/dc/terms/replaces
terms:replaces a owl:NamedIndividual .
# http://purl.org/dc/terms/requires
terms:requires a owl:NamedIndividual .
# http://purl.org/dc/terms/rights
terms:rights a owl:NamedIndividual .
# http://purl.org/dc/terms/rightsHolder
terms:rightsHolder a owl:NamedIndividual .
# http://purl.org/dc/terms/source
terms:source a owl:NamedIndividual .
# http://purl.org/dc/terms/spatial
terms:spatial a owl:NamedIndividual .
# http://purl.org/dc/terms/subject
terms:subject a owl:NamedIndividual .
# http://purl.org/dc/terms/tableOfContents
terms:tableOfContents a owl:NamedIndividual .
# http://purl.org/dc/terms/temporal
terms:temporal a owl:NamedIndividual .
# http://purl.org/dc/terms/title
terms:title a owl:NamedIndividual .
# http://purl.org/dc/terms/type
terms:type a owl:NamedIndividual .
# http://purl.org/dc/terms/valid
terms:valid a owl:NamedIndividual .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#decryptedStatus
nfo:decryptedStatus a owl:NamedIndividual ;
        rdfs:label "DecryptedStatus" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#encryptedStatus
nfo:encryptedStatus a owl:NamedIndividual ;
        rdfs:label "EncryptedStatus" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#hasHash
nfo:hasHash a owl:NamedIndividual .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#losslessCompressionType
nfo:losslessCompressionType a owl:NamedIndividual ;
        rdfs:label "losslessCompressionType" .
# http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#lossyCompressionType
nfo:lossyCompressionType a owl:NamedIndividual ;
        rdfs:label "lossyCompressionType" .
# #    Annotations
terms:Box terms:issued "2000-07-11"^^xsd:date ;
        rdfs:comment "The set of regions in space defined by their geographic coordinates according to the
DCMI Box Encoding Scheme."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:seeAlso <http://dublincore.org/documents/dcmi-box/> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "DCMI Box"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#Box-003> .
#
terms:ISO3166 terms:issued "2000-07-11"^^xsd:date ;
        rdfs:comment  "The  set  of  codes  listed  in  ISO  3166-1  for  the  representation  of  names  of
countries."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#ISO3166-004> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "ISO 3166"@en ;
        rdfs:seeAlso        <http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-
en1.html> .
```

```
#
terms:ISO639-2 rdfs:seeAlso <http://lcweb.loc.gov/standards/iso639-2/langhome.html> ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:comment "The three-letter alphabetic codes listed in ISO639-2 for the representation of names
of languages."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "ISO 639-2"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#ISO639-2-003> .
#
terms:ISO639-3 rdfs:label "ISO 639-3"@en ;
        rdfs:comment "The set of three-letter codes listed in ISO 639-3 for the representation of names of
languages."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#ISO639-3-001> ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:seeAlso <http://www.sil.org/iso639-3/> .
#
terms:Period rdfs:comment "The set of time intervals defined by their limits according to the DCMI Period
Encoding Scheme."@en ;
        rdfs:seeAlso <http://dublincore.org/documents/dcmi-period/> ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "DCMI Period"@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#Period-003> .
#
terms:Point rdfs:comment "The set of points in space defined by their geographic coordinates according to
the DCMI Point Encoding Scheme."@en ;
        rdfs:label "DCMI Point"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#Point-003> ;
        rdfs:seeAlso <http://dublincore.org/documents/dcmi-point/> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:Policy rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Policy"@en ;
        rdfs:comment "A plan or course of action by an authority, intended to influence and determine
decisions, actions, and other matters."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#Policy-001> ;
        terms:issued "2008-01-14"^^xsd:date .
#
terms:RFC1766 rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#RFC1766-003> ;
        rdfs:comment "The set of tags, constructed according to RFC 1766, for the identification of
languages."@en ;
        rdfs:seeAlso <http://www.ietf.org/rfc/rfc1766.txt> ;
        rdfs:label "RFC 1766"@en .
#
terms:RFC3066 terms:issued "2002-07-13"^^xsd:date ;
        rdfs:label "RFC 3066"@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:description "RFC 3066 has been obsoleted by RFC 4646."@en ;
        rdfs:comment "The set of tags constructed according to RFC 3066 for the identification of
languages."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:seeAlso <http://www.ietf.org/rfc/rfc3066.txt> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#RFC3066-002> .
#
terms:RFC4646 terms:description "RFC 4646 obsoletes RFC 3066."@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#RFC4646-001> ;
        rdfs:seeAlso <http://www.ietf.org/rfc/rfc4646.txt> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "The set of tags constructed according to RFC 4646 for the identification of
languages."@en ;
        rdfs:label "RFC 4646"@en .
#
terms:RFC5646 terms:issued "2010-10-11"^^xsd:date ;
        rdfs:seeAlso <http://www.ietf.org/rfc/rfc5646.txt> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#RFC5646-001> ;
        rdfs:comment "The set of tags constructed according to RFC 5646 for the identification of
languages."@en ;
        rdfs:label "RFC 5646"@en ;
        terms:description "RFC 5646 obsoletes RFC 4646."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:URI rdfs:comment "The set of identifiers constructed according to the generic syntax for Uniform
Resource Identifiers as specified by the Internet Engineering Task Force."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:seeAlso <http://www.ietf.org/rfc/rfc3986.txt> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#URI-003> ;
        rdfs:label "URI"@en ;
        terms:modified "2008-01-14"^^xsd:date .
```

```
#
terms:W3CDTF rdfs:label "W3C-DTF"@en ;
        rdfs:comment "The set of dates and times constructed according to the W3C Date and Time Formats
Specification."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#W3CDTF-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:seeAlso <http://www.w3.org/TR/NOTE-datetime> ;
        terms:modified "2008-01-14"^^xsd:date .
#
terms:abstract rdfs:label "Abstract"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#abstract-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:comment "A summary of the resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:accessRights terms:description "Access Rights may include information regarding access or
restrictions based on privacy, security, or other policies."@en ;
        terms:issued "2003-02-15"^^xsd:date ;
        rdfs:comment "Information about who can access the resource or an indication of its security
status."@en ;
        rdfs:label "Access Rights"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#accessRights-002> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:accrualMethod terms:hasVersion <http://dublincore.org/usage/terms/history/#accrualMethod-003> ;
        rdfs:label "Accrual Method"@en ;
        terms:modified "2010-10-11"^^xsd:date ;
        rdfs:comment "The method by which items are added to a collection."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2005-06-13"^^xsd:date .
#
terms:accrualPeriodicity terms:modified "2010-10-11"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#accrualPeriodicity-003> ;
        rdfs:label "Accrual Periodicity"@en ;
        rdfs:comment "The frequency with which items are added to a collection."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2005-06-13"^^xsd:date .
#
terms:accrualPolicy rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2005-06-13"^^xsd:date ;
        terms:modified "2010-10-11"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#accrualPolicy-003> ;
        rdfs:comment "The policy governing the addition of items to a collection."@en ;
        rdfs:label "Accrual Policy"@en .
#
terms:alternative rdfs:comment "An alternative name for the resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#alternative-003> ;
        terms:modified "2010-10-11"^^xsd:date ;
        terms:description "The distinction between titles and alternative titles is application-
specific."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Alternative Title"@en .
#
terms:audience terms:issued "2001-05-21"^^xsd:date ;
        rdfs:comment "A class of entity for whom the resource is intended or useful."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#audience-003> ;
        rdfs:label "Audience"@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:modified "2008-01-14"^^xsd:date .
#
terms:available rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "Date (often a range) that the resource became or will become available."@en ;
        rdfs:label "Date Available"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#available-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date .
#
terms:bibliographicCitation terms:modified "2008-01-14"^^xsd:date ;
        terms:description "Recommended practice is to include sufficient bibliographic detail to identify
the resource as unambiguously as possible."@en ;
        rdfs:label "Bibliographic Citation"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#bibliographicCitation-002> ;
        rdfs:comment "A bibliographic reference for the resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2003-02-15"^^xsd:date .
#
terms:conformsTo rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Conforms To"@en ;
        rdfs:comment "An established standard to which the described resource conforms."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#conformsTo-003> ;
        terms:issued "2001-05-21"^^xsd:date .
#
terms:contributor rdfs:comment "An entity responsible for making contributions to the resource."@en ;
```

```
        terms:modified "2010-10-11"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:description "Examples of a Contributor include a person, an organization, or a service."@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#contributorT-001> ;
        rdfs:label "Contributor"@en .
#
terms:coverage rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#coverageT-001> ;
        terms:description "Spatial topic and spatial applicability may be a named place or a location
specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A
jurisdiction may be a named administrative entity or a geographic place to which the resource applies.
Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names
[TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers
such as sets of coordinates or date ranges."@en ;
        rdfs:label "Coverage"@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:comment "The spatial or temporal topic of the resource, the spatial applicability of the
resource, or the jurisdiction under which the resource is relevant."@en ;
        terms:modified "2008-01-14"^^xsd:date .
#
terms:created rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "Date of creation of the resource."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#created-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "Date Created"@en ;
        terms:issued "2000-07-11"^^xsd:date .
#
terms:creator rdfs:comment "An entity primarily responsible for making the resource."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#creatorT-002> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:modified "2010-10-11"^^xsd:date ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:label "Creator"@en ;
        terms:description "Examples of a Creator include a person, an organization, or a service."@en .
#
terms:date rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "A point or period of time associated with an event in the lifecycle of the
resource."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#dateT-001> ;
        terms:description "Date may be used to express temporal information at any level of granularity.
Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601
[W3CDTF]."@en ;
        rdfs:label "Date"@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date .
#
terms:dateAccepted terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#dateAccepted-002> ;
        terms:issued "2002-07-13"^^xsd:date ;
        terms:description "Examples of resources to which a Date Accepted may be relevant are a thesis
(accepted by a university department) or an article (accepted by a journal)."@en ;
        rdfs:label "Date Accepted"@en ;
        rdfs:comment "Date of acceptance of the resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:dateCopyrighted rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Date Copyrighted"@en ;
        rdfs:comment "Date of copyright."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#dateCopyrighted-002> ;
        terms:issued "2002-07-13"^^xsd:date .
#
terms:dateSubmitted terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "Date of submission of the resource."@en ;
        terms:description "Examples of resources to which a Date Submitted may be relevant are a thesis
(submitted to a university department) or an article (submitted to a journal)."@en ;
        terms:issued "2002-07-13"^^xsd:date ;
        rdfs:label "Date Submitted"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#dateSubmitted-002> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:description terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "Description"@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#descriptionT-001> ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:comment "An account of the resource."@en ;
        terms:description "Description may include but is not limited to: an abstract, a table of contents,
a graphical representation, or a free-text account of the resource."@en .
#
terms:educationLevel terms:modified "2008-01-14"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2002-07-13"^^xsd:date ;
        rdfs:label "Audience Education Level"@en ;
        rdfs:comment "A class of entity, defined in terms of progression through an educational or training
context, for which the described resource is intended."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#educationLevel-002> .
```

```
#
terms:extent terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "The size or duration of the resource."@en ;
        rdfs:label "Extent"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#extent-003> .
#
terms:format terms:issued "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#formatT-001> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "The file format, physical medium, or dimensions of the resource."@en ;
        terms:description "Examples of dimensions include size and duration. Recommended best practice is
to use a controlled vocabulary such as the list of Internet Media Types [MIME]."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Format"@en .
#
terms:hasFormat rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#hasFormat-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "Has Format"@en ;
        rdfs:comment "A related resource that is substantially the same as the pre-existing described
resource, but in another format."@en ;
        terms:issued "2000-07-11"^^xsd:date .
#
terms:hasPart rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Has Part"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "A related resource that is included either physically or logically in the described
resource."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#hasPart-003> .
#
terms:hasVersion terms:hasVersion <http://dublincore.org/usage/terms/history/#hasVersion-003> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "A related resource that is a version, edition, or adaptation of the described
resource."@en ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        rdfs:label "Has Version"@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date .
#
terms:identifier terms:hasVersion <http://dublincore.org/usage/terms/history/#identifierT-001> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:description "Recommended best practice is to identify the resource by means of a string
conforming to a formal identification system. "@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "An unambiguous reference to the resource within a given context."@en ;
        rdfs:label "Identifier"@en ;
        terms:issued "2008-01-14"^^xsd:date .
#
terms:instructionalMethod terms:issued "2005-06-13"^^xsd:date ;
        terms:description "Instructional Method will typically include ways of presenting instructional
materials or conducting instructional activities, patterns of learner-to-learner and learner-to-instructor
interactions, and mechanisms by which group and individual levels of learning are measured.  Instructional
methods include all aspects of the instruction and learning processes from planning and implementation
through evaluation and feedback."@en ;
        rdfs:label "Instructional Method"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "A process, used to engender knowledge, attitudes and skills, that the described
resource is designed to support."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#instructionalMethod-002> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:isFormatOf terms:hasVersion <http://dublincore.org/usage/terms/history/#isFormatOf-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "A related resource that is substantially the same as the described resource, but in
another format."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        rdfs:label "Is Format Of"@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:isPartOf terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "Is Part Of"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#isPartOf-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "A related resource in which the described resource is physically or logically
included."@en .
#
terms:isReferencedBy rdfs:label "Is Referenced By"@en ;
        rdfs:comment "A related resource that references, cites, or otherwise points to the described
resource."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#isReferencedBy-003> .
#
terms:isReplacedBy rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Is Replaced By"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "A related resource that supplants, displaces, or supersedes the described
resource."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#isReplacedBy-003> .
#
terms:isRequiredBy rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Is Required By"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#isRequiredBy-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "A related resource that requires the described resource to support its function,
delivery, or coherence."@en .
#
terms:isVersionOf rdfs:label "Is Version Of"@en ;
        rdfs:comment "A related resource of which the described resource is a version, edition, or
adaptation."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2000-07-11"^^xsd:date ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#isVersionOf-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:description "Changes in version imply substantive changes in content rather than differences
in format."@en .
#
terms:issued terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "Date of formal issuance (e.g., publication) of the resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#issued-003> ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Date Issued"@en .
#
terms:language rdfs:label "Language"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#languageT-001> ;
        terms:issued "2008-01-14"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:description "Recommended best practice is to use a controlled vocabulary such as RFC 4646
[RFC4646]."@en ;
        rdfs:comment "A language of the resource."@en .
#
terms:license terms:issued "2004-06-14"^^xsd:date ;
        rdfs:label "License"@en ;
        rdfs:comment "A legal document giving official permission to do something with the resource."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#license-002> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:mediator terms:description "In an educational context, a mediator might be a parent, teacher,
teaching assistant, or care-giver."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "An entity that mediates access to the resource and for whom the resource is intended
or useful."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#mediator-003> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Mediator"@en ;
        terms:issued "2001-05-21"^^xsd:date .
#
terms:medium rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Medium"@en ;
        rdfs:comment "The material or physical carrier of the resource."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#medium-003> .
#
terms:modified rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Date Modified"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#modified-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "Date on which the resource was changed."@en ;
        terms:issued "2000-07-11"^^xsd:date .
#
terms:provenance rdfs:comment "A statement of any changes in ownership and custody of the resource since
its creation that are significant for its authenticity, integrity, and interpretation."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2004-09-20"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:description "The statement may include a description of any changes successive custodians
made to the resource."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#provenance-002> ;
        rdfs:label "Provenance"@en .
#
terms:publisher terms:modified "2010-10-11"^^xsd:date ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:label "Publisher"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#publisherT-001> ;
        terms:description "Examples of a Publisher include a person, an organization, or a service."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "An entity responsible for making the resource available."@en .
#
terms:references rdfs:comment "A related resource that is referenced, cited, or otherwise pointed to by the
described resource."@en ;
        rdfs:label "References"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#references-003> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en .
#
terms:relation terms:issued "2008-01-14"^^xsd:date ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#relationT-001> ;
        rdfs:label "Relation"@en ;
        terms:description "Recommended best practice is to identify the related resource by means of a
string conforming to a formal identification system. "@en ;
        rdfs:comment "A related resource."@en ;
        terms:modified "2008-01-14"^^xsd:date .
#
terms:replaces skos:note "This term is intended to be used with non-literal values as defined in the DCMI
Abstract Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage
Board is seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "Replaces"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#replaces-003> ;
        rdfs:comment "A related resource that is supplanted, displaced, or superseded by the described
resource."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:requires terms:hasVersion <http://dublincore.org/usage/terms/history/#requires-003> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:label "Requires"@en ;
        rdfs:comment "A related resource that is required by the described resource to support its
function, delivery, or coherence."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2000-07-11"^^xsd:date .
#
terms:rights rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:label "Rights"@en ;
        rdfs:comment "Information about rights held in and over the resource."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#rightsT-001> ;
        terms:description "Typically, rights information includes a statement about various property rights
associated with the resource, including intellectual property rights."@en .
#
terms:rightsHolder rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#rightsHolder-002> ;
        terms:modified "2008-01-14"^^xsd:date ;
        rdfs:comment "A person or organization owning or managing rights over the resource."@en ;
        terms:issued "2004-06-14"^^xsd:date ;
        rdfs:label "Rights Holder"@en .
#
terms:source rdfs:label "Source"@en ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#sourceT-001> ;
        rdfs:comment "A related resource from which the described resource is derived."@en ;
        terms:description "The described resource may be derived from the related resource in whole or in
part. Recommended best practice is to identify the related resource by means of a string conforming to a
formal identification system."@en .
#
terms:spatial terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Spatial Coverage"@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:comment "Spatial characteristics of the resource."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#spatial-003> .
#
terms:subject terms:modified "2012-06-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#subjectT-002> ;
        skos:note "This term is intended to be used with non-literal values as defined in the DCMI Abstract
Model (http://dublincore.org/documents/abstract-model/).  As of December 2007, the DCMI Usage Board is
seeking a way to express this intention with a formal range declaration."@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        rdfs:label "Subject"@en ;
        rdfs:comment "The topic of the resource."@en ;
        terms:description "Typically, the subject will be represented using keywords, key phrases, or
classification codes. Recommended best practice is to use a controlled vocabulary."@en ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:tableOfContents rdfs:comment "A list of subunits of the resource."@en ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        rdfs:label "Table Of Contents"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#tableOfContents-003> .
#
terms:temporal terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Temporal Coverage"@en ;
        rdfs:comment "Temporal characteristics of the resource."@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#temporal-003> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
terms:title rdfs:label "Title"@en ;
        rdfs:comment "A name given to the resource."@en ;
        terms:issued "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#titleT-002> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:modified "2010-10-11"^^xsd:date .
#
terms:type rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
        terms:description "Recommended best practice is to use a controlled vocabulary such as the DCMI
Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource,
use the Format element."@en ;
        rdfs:comment "The nature or genre of the resource."@en ;
        rdfs:label "Type"@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#typeT-001> ;
        terms:issued "2008-01-14"^^xsd:date .
#
terms:valid rdfs:comment "Date (often a range) of validity of a resource."@en ;
        terms:modified "2008-01-14"^^xsd:date ;
        terms:issued "2000-07-11"^^xsd:date ;
        rdfs:label "Date Valid"@en ;
        terms:hasVersion <http://dublincore.org/usage/terms/history/#valid-003> ;
        rdfs:isDefinedBy <http://purl.org/dc/terms/> .
#
:hasDescription rdfs:label "hasDescription"@en .
#
:hasIdentifier rdfs:label "hasIdentifier"@en .
#
:hasTitle rdfs:label "hasTitle"@en .
#
nfo:belongsToContainer rdfs:comment "Models the containment relations between Files and Folders (or
CompressedFiles)." ;
        rdfs:label "belongsToContainer" .
#
nfo:bookmarks rdfs:comment "The address of the linked object. Usually a web URI." ;
        nrl:cardinality "1"^^xsd:integer ;
        rdfs:label "link" .
#
nfo:conflicts rdfs:label "conflicts" ;
```

```
        rdfs:comment "States that a piece of software is in conflict with another piece of software." .
#
nfo:containsBookmark rdfs:label "contains bookmark" ;
        rdfs:comment "The folder contains a bookmark." .
#
nfo:containsBookmarkFolder rdfs:comment "The folder contains a bookmark folder." ;
        rdfs:label "contains folder" .
#
nfo:containsPlacemark rdfs:label "contains Placemark" ;
        rdfs:comment "Containment relation between placemark containers (files) and placemarks within." .
#
nfo:fileCreated rdfs:comment "File creation date" ;
        rdfs:label "fileCreated" ;
        nrl:maxCardinality "1"^^xsd:integer .
#
nfo:fileLastAccessed rdfs:comment "Time when the file was last accessed." ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:label "fileLastAccessed" .
#
nfo:fileLastModified nao:deprecated "true"^^xsd:boolean ;
        rdfs:comment "last modification date" ;
        rdfs:label "fileLastModified" .
#
nfo:fileName rdfs:comment "Name of the file, together with the extension" ;
        rdfs:label "fileName" ;
        nrl:maxCardinality "1" .
#
nfo:fileSize nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:label "fileSize" ;
        rdfs:comment "The size of the file in bytes. For compressed files it means the size of the packed
file, not of the contents. For folders it means the aggregated size of all contained files and folders " .
#
nfo:fileUrl nao:deprecated "true"^^xsd:boolean ;
        rdfs:label "fileUrl" ;
        rdfs:comment "URL of the file. It points at the location of the file. In cases where creating a
simple file:// or http:// URL for a file is difficult (e.g. for files inside compressed archives) the
applications are encouraged to use conventions defined by Apache Commons VFS Project at
http://jakarta.apache.org/  commons/ vfs/ filesystems.html." .
#
nfo:foundry rdfs:label "foundry" ;
        nrl:maxCardinality "1"^^xsd:integer ;
        rdfs:comment "The foundry, the organization that created the font." .
#
nfo:hasHash rdfs:comment "Links the file with it's hash value." ;
        nao:userVisible "false"^^xsd:boolean ;
        rdfs:label "hasHash" .
#
nfo:hasMediaStream rdfs:label "hasMediaStream" ;
        rdfs:comment "Connects a media container with a single media stream contained within." .
#
nfo:supercedes rdfs:comment "States that a piece of software supercedes another piece of software." ;
        rdfs:label "supercedes" .
#
nfo:uuid rdfs:label "uuid" ;
        rdfs:comment "Universally unique identifier of the filesystem. In the future, this property may
have its parent changed to a more generic class." .
```

Documentation

The full source code for the COSI Ontology is documented in FigShare (Brahaj, 2015). A description and relation of all classes and properties is listed below.

Acknowledgements

I extend my sincere gratitude and appreciation to many people who helped me complete this work. The first person I would like to thank is my mentor Matthias Razum. This dissertation is highly motivated by his work and his guidance has been crucial to my progress. Together with Dr. Leni Helmes, Matthias made sure I can allocate the necessary time for my research in these years. I owe a lot of thanks to all my colleges of the FIZ Karlsruhe. Many thanks to Frank Schwichtenberg, a great friend and colleague who made sure I am never overloaded. Many thanks to Stefan Bozic, Christian Herlambang and other colleagues at FIZ who supported my work and helped me face my challenges.

I wish to thank very much Dr. Peter Schirmbacher who accepted to be my supervisor and a wise guide in this journey. I owe you lots of gratitude for the patience and trust you put in me. Dr. Vivien Petras who picked me halfway and helped me make it to the end of the journey. I am grateful to her for the ingenious insights and feedback that helped me improve my work. And last but not least, my research would have never started without the stimulus by Dr. Detlev Doherr with whom I collaborated since my master thesis and convinced me I should go further. I will always remember discussions we had and the way you inspired and transmitted enthusiasm in us.

Finally, I owe a lot of thanks to my family. My wonderful wife Julia, for she is my muse and source of strength. We shared a lot of burden together; both started the doctoral journey on the same time but we managed to work like a perfect team and succeed. My daughter Emma for being that star guiding me into peaceful (non-research) moments every evening. You brought joy and colours to every day of my life. My son Bato, for he just saw the word and he has no idea what I did in the past years. My parents who valued every little victory I had and taught me to never give up. You guys are my life!

Look Mom, I did it!

Declaration

Hiermit erkläre ich, Armand BRAHAJ Matrikel-Nr: 532011, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Dissertation wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt oder veröffentlicht.

Karlsruhe, den 01.06.2016

Armand Brahaj