TECHNICAL REPORT

LONGREC

PRESERVATION OF SEMANTIC VALUE - A STUDY OF THE STATE OF THE ART

REPORT NO 2008-0123

REVISION NO 1.01

DET NORSKE VERITAS

TECHNICAL REPORT

Date of first issue: 13.11.2007	Project No: 913A0304
Approved by: Process and information technolgy Head of program	Organisational unit: BRINO0913
Client: LongRec	Client ref.:

DET NORSKE VERITAS AS

1322 Høvik
Norway
Tel:
Fax:
http://www.dnv.com
NO 945 748 931 MVA

Summary:

The purpose of this state of the art within digital information preservation is to look into the fundamental problem that it is difficult to assure correct interpretation of the content in an old record. This problem is due to changes in a set of factors, some of these factors are: (i) the symbols/words used, (ii) the meaning of symbols/words and (iii) the domain knowledge. An extensive literature study has been performed into areas such as records preservation, library science, literature science, semantic technology and others. Literature on these topics is somewhat limited, and we therefore believe that this is a somewhat immature research area.

Areas relevant to change of semantic value in records are identified, and six of these areas (symbol, reference, referent, record, worldview and countermeasures) are described and discussed in this report, together with a review of current best practice.

Report No: 2008-0123	Subject Group:	
Report title: Preservation of semantic value - a study of the state of the art		
Work carried out by: Per Myrseth, Tore R. Christiansen, Helen Gayorfar		
Work verified by:		
Date of this revision: 1. April 2008	Revision No: 1.01	Number of pages: 28

Indexing terms

Keywords	Service Area
symbol, reference, referent, record, worldview and countermeasures, semantic technology, semantic preservation, semantic value	
	Market Sector

☒ Unrestricted distribution (internal and external)

□ Unrestricted distribution within DNV

□ Limited distribution within DNV after 3 years

□ No distribution (confidential)

Table of Contents *Page*

1 INTRODUCTION

1.1 Executive summary

The purpose of this state of the art within digital information preservation is to look into the fundamental problem that it is difficult to assure correct interpretation of the content in an old record. This problem is due to changes in a set of factors, some of these factors are: (i) the symbols/words used, (ii) the meaning of symbols/words and (iii) the domain knowledge. An extensive literature study has been performed into areas such as records preservation, library science, literature science, semantic technology and others. Literature on preservation of semantic value is found to be somewhat limited, reflecting the state of a relatively immature research area.

Areas relevant to change of the semantic value in records (or parts thereof) are identified, and six of these areas (symbol/words, reference/thought, referent/object, records in it self, worldview and countermeasures) are described and discussed in this report, together with a review of current best practice.

1.2 Background

The objective of this report is to describe state of the art in technologies, methodologies and research related to the preservation of semantic value in records. In this report we use the term record as an alias to the term “Information Package” from Open Archival Information System Reference Model (OAIS) [35]**.** I.e. a record is a digital container for data of any type, e.g. text, pictures, 3D-drawings, sound, video supplemented with its preservation metadata.

Based on experience from linguistics and studies of development in languages, we believe a good regime for preservation of digital records must include solutions to challenges also outside the IT-domain. It can be difficult for people to understand old records precisely, and even worse for computers to interpret and analyse their content and meaning correctly.

Understanding is a sliding scale, a user may understand little, some or almost everything of a record, but not all. From a preservation of records point of view, it is a challenge whether or not a user of a record understand and interprets the writers intention.

The meaning of data and their intended use may change and evolve over time. As time passes there will be changes in knowledge, jurisdiction and practice, use of language, culture, society at large, software, ontologies[1], pattern of how software and processes use the records and their corresponding ontologies. Even innovation in seemingly non-related areas could affect the meaning of data.

This report is a part of the LongRec (Long-Term Records Management) project run by Det Norske Veritas (DNV) in collaboration with a number of case partners, commercialization partners and research partners. The primary objective of LongRec is the persistent, reliable and trustworthy long-term archival of digital information records with emphasis on availability and use of the information. The project’s public web site is at http://research.dnv.com/longrec/

[1] Simplified an ontology is a set of references, their definition, attributes and areas of use.

LongRec is a three year project (2007-2009) partly funded by the Norwegian Research Council. The project constitutes the Norwegian team of the InterPARES 3 project, http://www.interpares.org. LongRec addresses several research challenges, each of which is assigned a short name (in parentheses below): records transition survival (READ), long-term usage (FIND), preservation of semantic value (UNDERSTAND), preservation of evidential value (TRUST) and legal, social, and cultural framework (COMPLIANCE).

1.3 Scope of this report

The scope of this report is preservation of semantic value in records. Since records most often contain and represent information and knowledge about an object (thing, activity, period, state, etc.), the scope must also focus on the understanding of the object itself, its use and the context in which it was created and intended to be used.

The project scope of LongRec defines a set of prerequisites for being able to use records over time. These are:

- Finding the record
- Reading the record
- Presenting/viewing/rendering the record
- Trusting the record

However, can the record be useful if you don't understand it?

In the LongRec UNDERSTAND research area we look into changes that over time have a semantic impact. The *areas of change* / areas that have a lifecycle are:

1. How records are used/procedures (including reasoning about content)
2. Laws and regulations
3. Users of the records
4. Presentation of records
5. **Symbol**/terminology related to the actual topic/concept
6. **Reference**/thought, i.e. knowledge and meaning about the actual topic (ontology)
7. **Referent**/ the object itself (a ship versus a record describing a ship)
8. **Record**/data about an object/referent
9. **Worldview** and context
10. **Countermeasures**, are used to slow down the deterioration of semantics and/or log the changes in semantics to later be able to inform coming users of related records. They are used to reduce the semantic consequence of the other *areas of change* listed above. But the countermeasures in it self changes. The two main types of countermeasures are:
    - Procedural and administrative countermeasures
    - Technology based countermeasures

(Text in bold in the bullets above are discussed in separate chapters later in this report)

In this report we focus on state of the art related to *areas of change* listed 5 to 10 above. The points numbered 1-4 are at the time being, regarded as out of scope for this state of the art. Some or all of these may however be important to the overall objective of LongRec and may be important to the case studies of LongRec.

An important source for this report is the reflections and the state of the art described in [1]. The article is from the cultural and library field, and has a focus on handling semantics through changes in time, context, worldviews, needs and technology. We have also searched literature in the areas linguistics, computer science, ontology, metadata and conservation.

1.4 How to read this report

The structure of this report is based on the semantic triangle from Ogden [3], which is described in chapter 2, Breakdown of the Understand topics. The next three chapters describe the corners in the triangle: Symbol, referent and reference/ontology. Next is chapter 6, Record, describing aspects of the content to be preserved. Chapter 7 introduces worldviews and context, and finally chapter 8, Countermeasures discusses means used to reduce the consequence of loss of semantic value in the *areas of change*. Figure 1, Chapter illustration, illustrates, based on the semantic triangle of Ogden [3], how most of the chapters are linked.

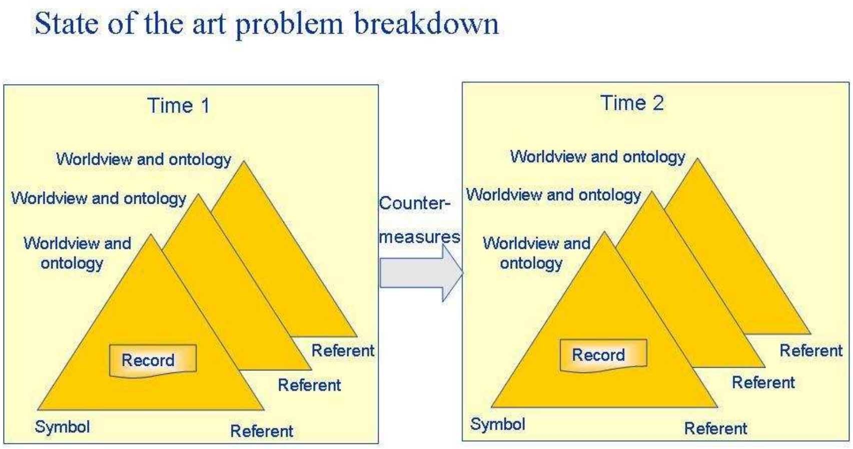
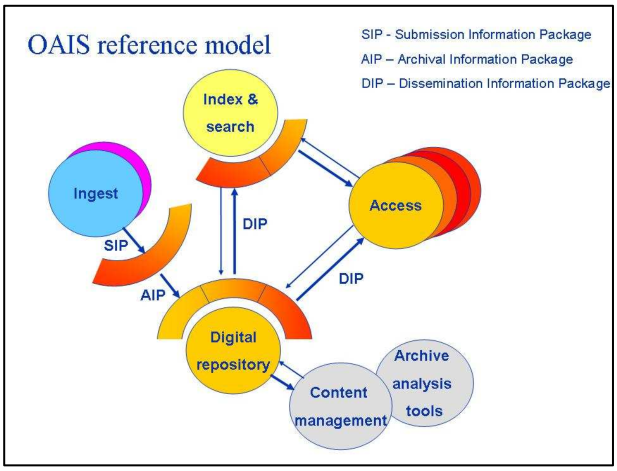


Figure 1, Chapter illustration

Each of the terms in the figure above is described in separate chapters later in this report. The audience of this report is the project participants, InterPARES and academia working with semantic technologies and content management systems.



In the archival domain metadata has traditionally been used in various ways to handle aspects of preservation. Since metadata is such a broad topic in itself, the reader will find that chapters about semantic annotation, ontology engineering etc. will cover the topic that in archival terminology would have been called metadata. Based on the understanding of OAIS reference model [35] shown in Figure 2, OAIS Reference Model, we will in this report drill into areas of this architecture where the semantics of the records are captured, maintained, changed and improved.

Figure 2, OAIS Reference Model

1.5 History of semantics

The history of semantics leads us back to the ancient Greece where the study of words started two and a half thousand years ago. The work of Socrates, Plato and Aristotle is the historical starting point for the development leading to the concepts and technologies of the Semantic Web.
Two important books on this globe is the Bible and the Koran. From a LongRec Understand work package view, they both raise several semantic challenges, and they are probably the most discussed books ever.

Historically man started with oral communication and supplemented it with carvings, icons, runes and later letters. After the introduction of TV, computer screens, cheaper printing and colour printing, mobile phones etc. symbols, icons and logos are becoming even more important. Examples are traffic signs, logos for commercial brands, icons on remote controls and mobile phones, front panel on technical devices. With increased globalisation the contact between cultures with different backgrounds / worldviews raise the topics of semantics to new heights.

Another example is the save icon in Microsoft Windows applications showing a diskette. Since floppy-disks are an outdated technology and in rapidly decreasing use, it is interesting to see what icon will replace the floppy-disk. An extreme examples is the challenge of describing to coming generations principles for the long term storage of nuclear waste for 100 000 years. This description may require some form of language independent symbolism.

1.6 Best practice

To help the reader understand the need for the LongRec Understand work package, an introduction on best practice will be useful background knowledge.

Records preserved in public archival and library organisations are supplemented with a set of context metadata for capturing classification, author, relevant dates etc. In [28] after a discussion on the importance of using record formats for long term storage, a reflection of semantics is stated as follows: "*in order to understand the document (record), it is equally important to capture the context of the document (record), generally as metadata.*" This use of metadata is a common approach in preservation and data quality regimes.

The term context metadata does not cover the *areas of change* listed in the scope chapter of this report. We believe that this tradition of context metadata alone is not sufficient to preserve the semantic value.

It is common that each organisation has its own local archive and after some period (10 years or more) the records should be transferred to a preservation repository at e.g. The National Library of Norway or The National Archival Service of Norway.

The Brønnøysund Register Centres "Register of Business Enterprises" is the authoritative source for Norwegian legal entities holding information like company number, name, purpose, board members for every Norwegian company. This is a transaction based system where the latest version of the data is of most interest. All changes in data are kept in a history change log, so the information valid at a certain time can be found.

2 BREAKDOWN OF THE UNDERSTAND TOPICS

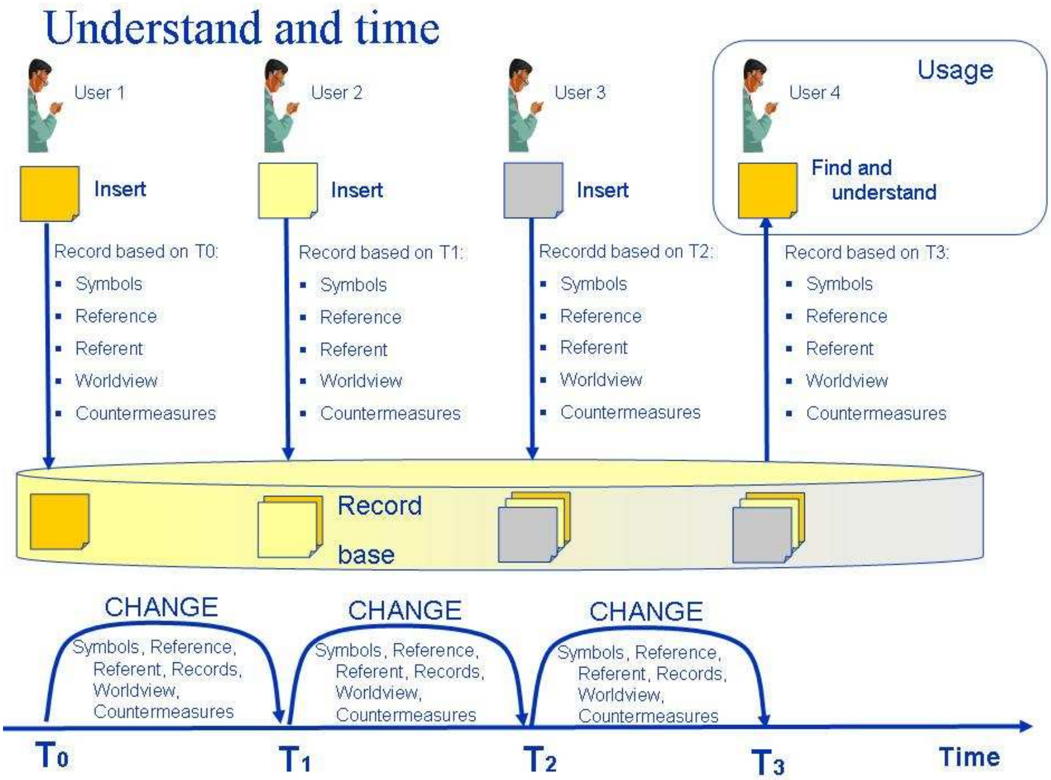


Figure 3, record usages and time

Figure 3, record usages and time, illustrates a timeline and submissions of records to a repository at different points in time. Later the submitted records are to be found, read, trusted and used, and the need to understand the content arises. The user at time T3 needs to be aware of the changes in knowledge base, context, term/symbol, usage, rules and regulations etc. otherwise the understanding and use of old records may lead to wrong conclusions, and merged data may be misleading. Figure 3 try to illustrate the lifecycle of each of the areas of change, i.e. the symbols, reference, referent, record, worldview and countermeasures between T0, T1, T2 and T3. At T3 the challenge of understanding and using each of these areas of change arises. Harvesting and use of different types of metadata will play a crucial role in coping with the different type of changes.

.

On basis of the limitation of scope for this state of the art, presented in chapter 1.3, we need to look into existing work in the area. Based on the work of Ogden [3], ISO 1087 [12] and Veltman [1] we have identified several relevant elements to look into.

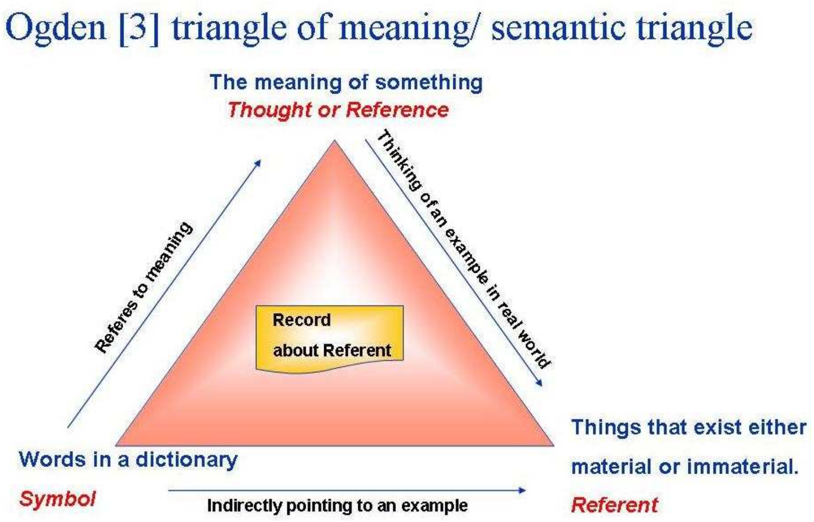


Figure 4, the relations between things, words and meaning[2]. Captions of the corners in red color and italic is same as the captions used in Ogden [3].

The records or data are placed in the middle of the triangle illustrating that the records use symbols, and that the symbols have as link to reference/meaning. Further, the records are about referents. The book of Ogden, Meaning of Meaning [3] set out principles for understanding the function of language, and describes the so-called semantic triangle[3]. The link between symbols and referents are of great importance and the source to many challenges. This challenge is mentioned as follows in [3, p12]:

> *We shall find, however, that the kind of simplification typified by this once universal theory of direct meaning relations between words (symbols) and things (referents) is the source of almost all the difficulties which thought encounters. As will appear at a later stage, the power to confuse and obstruct, which such simplifications possess, is largely due to the conditions of communication.*

The semantic triangle illustrates a recursive challenge similar to what we have when we do ontology modelling, i.e. the description of a reference will use symbols pointing to other references as part of its definition. Thereby many instances of the triangle linked together will be the result if you try to make a complete model, where a symbol used in one triangle instance, will refer to another triangle instance.

In the context of information modelling and semantics, the relation between the triangle and OMGs levels of metamodels (MOF [42]) we find that

- Symbol and Referent do not have a direct relation
- Records relates to M0 data
- Reference relates to M1 model
- M2 do not have a direct relation to the triangle.

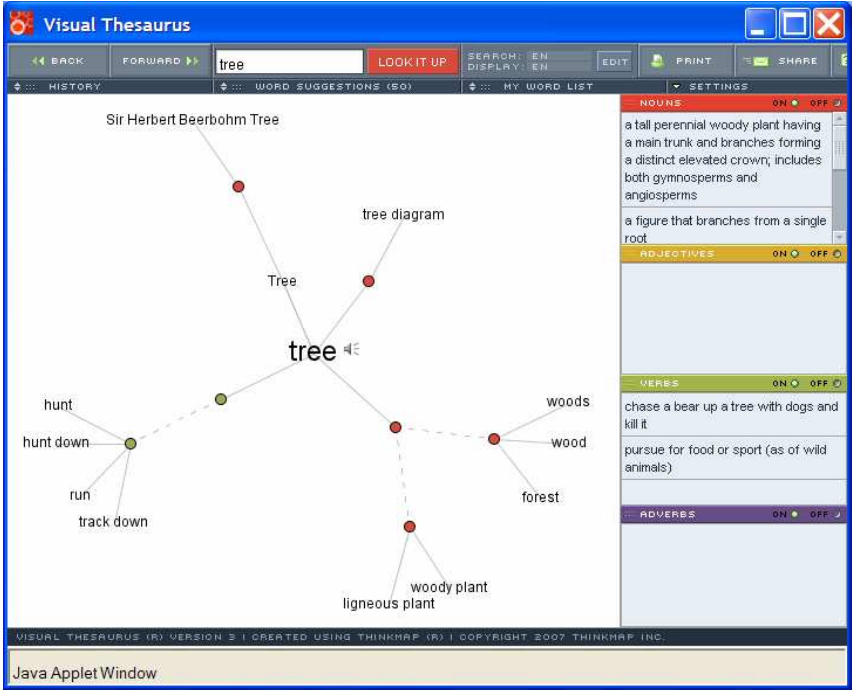
Below is the description and motivation for each of the chosen areas of change listed.

- **Symbol**: Is a word, an icon or a pattern used for identifying a reference, e.g. the word “tree”. A symbol is used to point to one or more references. The figure below shows different meanings

[2] Example of an immaterial referent could be house insurance. The insurance document may have a digital or paper based representation, but the insurance itself is immaterial.

[3] http://en.wikipedia.org/wiki/Charles_Kay_Ogden

(pointers to different references) of the word “tree”.



a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms

a figure that branches from a single root

chase a bear up a tree with dogs and kill it

pursue for food or sport (as of wild animals)

VISUAL THESAURUS (R) VERSION 3 | CREATED USING THINKMAP (R) | COPYRIGHT 2007 THINKMAP INC.

Java Applet Window

Figure 5 visualizing relations of words

- **Reference (or thought)**: Is a concept or class listing the common attributes of two or more referents. E.g. the reference of a tree describes what it takes to be a tree, and often shows which broader / more generic reference it is related to. One of the definitions of the symbol “tree” is defined in Webster online as: “a tall perennial woody plant having a main trunk…” [4]
- **Referent (object)**: Is an instance of a reference, an object. E.g. the specific tree I have in my garden.
- **Records about a referent**: Are text, pictures, numbers, videos, models etc. describing some aspects of a referent. The data/record itself contains symbols. E.g. a record related to the tree I have in my garden could describe what type of tree it is, when it was planted, who planted it, how it was planted, what is the need for yearly maintenance, how it should be treated to have a long life etc.

The semantic value of content in each corner of the triangle, i.e. symbol, reference and referent, will change as time passes. There are several reasons for this that could shortly be described as follows:

- Symbol: The words used in a language change as the people and the culture change. The use of some symbols fades out, new ones come in, and the meaning of some symbols change.
- Reference: The attributes of a reference change. E.g. what it takes to be a fast ferry changes as powerful engines and building materials develop. Similarly, the attributes of a a mobile phone 20 years ago are very different from today.
- Referent: A referent has a lifespan and the attributes relevant for the referent change. Often a referent changes reference as well. E.g. a human being is always a human being, but it starts as

[4] Merriam Webster http://www.m-w.com/dictionary/tree

a new born, continues with becoming youth, middle age, elderly etc. If a person gets ill, she/he will also be a patient for a while.

- Data/record about a referent: Since a referent has a lifespan, the record will often describe an aspect of a referent at a certain stage in time.

Figure 4, the relations between things, words and meaning. Captions of the corners in red color and italic is same as the captions used in Ogden [3]., focuses on the four elements, symbol, reference, referent and records as time passes by. This means that different worldviews at a certain point in time are not depicted in the figure.

Different types of worldviews at the same point in time are for example based on:

- Purpose of a concept (its function). The purpose of a referent may differ depending on the context and the process it is a part of. E.g. a heavy lexicon book may be used for pressing leaves in a herbarium, but the lexicon is not originally made for that purpose. A computer is a kind of multi-tool, even an effective criminal tool for Internet based crime.
- Where a referent is (space). E.g. in international food transport the location of the food has an influence on what information is needed to handle it properly and to avoid damage. This is because differences in temperature, humidity, bugs, salt water exposure during transportation affect the food quality. This means that handling the same type of food in Egypt and at Svalbard could be quite different.
- Cultural and religious context. E.g. whether you eat pork meat or not is dependent on different religious traditions and how the pig is slaughtered.
- Legal context. How one type of food is made and dealt with differs in different markets based on legislations, e.g. what drugs or chemicals one may use differs based on what legislation one must follow.

The motivation for including worldviews as a supplementary dimension is described in chapter 7.

Countermeasures are used to reduce the semantic consequence of the other *areas of change* listed above. But the countermeasures in it self changes. The two main types of countermeasures are (i) procedural and administrative countermeasures, and (ii) technology based countermeasures.

Based on the above breakdown structure this report will try to describe the state of the art on preservation of semantics structured by the following areas of change:

- Symbol
- Reference
- Referent
- Records about a referent
- Multiple worldviews co-existing
- Countermeasures

2.1 Case study, Understand

We plan to link these areas of change to how The Brønnøysund Register Center establish and maintain their: records of Norwegian companies, models of these records, historical versions of records and models. Further the users' different worldviews, and how the different worldviews influence how records are understood and interpreted are of great interest.

2.2 Relation between LongRec Understand and Find work packages

LongRec Find work package and state of the art report focus on search methodologies and related technology. Parts of the Find functions use technologies from the semantic technology field. Use of semantic technologies is related to e.g. use of ontologies or taxonomies for better indexing, comparing records or in the process of rating search results. New dynamic results clustering and query expansion by (synonym) dictionary are some of the new properties of advanced search technology [37] [5]. A trend is that search technology uses many techniques for automatic metadata harvesting.

Analyses of portal- and search-logs can be used as input for ontology evolution. According to [37, p10] "*enterprise search system provides tools, lists, libraries, or other administrative functions to support the process of classification itself, maintaining taxonomies, extracting ontologies from indexed content, or some other value-added process*".

Based on a repository of records, search technology can establish a list of all used symbols/words, and to some extent find relations between words by use of natural language techniques. Technology in the intersection of search and semantics is a valuable supplement and useful tools for the ontology engineering and evolution processes. Search technology in itself will not be able to replace an editorial process for establishing consensus ontology to be used within a domain. On the other hand, it is relevant to discuss how much effort we can afford to put into the work of establishing a consensus ontology, and what properties such an ontology should have.

In the LongRec Find Case study at The National Library of Norway, harvesting and use of metadata are exemplified, and many different types of metadata can be seen in action both at time of record archiving, conversion, search and presentation.

The intersection of Understand and Find is of great interest for LongRec. References [37] and [36] disagree whether there is focus on the intersection. According to [36] little efforts are put into innovative use of combining semantic- and search-technology.

[5] Examples of automatic clustering of search results see e.g. clusty.com (previous www.vivisimo.com): http://clusty.com/search?query=search%20technology&tb=vivi-transition&

3 SYMBOL

Symbols are objects, pictures, or other concrete representations of ideas, concepts, or other abstractions[6]. A symbol could be a word in a language, characters, digits, a traffic sign or an icon on your computer desktop.

3.1 Evolution of words and their use

For centuries dictionaries have been the main mechanism for maintaining and distributing the common use of words. New initiatives in linguistic ontologies like WordNet[7], and even e-commerce initiatives like UN standard products and services codes, UNSPSC[8], are listed as examples in [26].

Further, online services like www.wikipedia.org and Merriam-Webster online[9] are used for distributing common use, but they differ in the editorial process. Wikipedia and Wiktionary are authored by its users; this is different to Merriam-Webster, which is based on an editor function.

The history of dictionary development has taught us that the efforts of making one single comprehensive dictionary in the 19th century ended up with the recognition that the effort will not succeed. The real challenge lies in creating bridges in various ways between dictionaries.

> *Cultural terms have local, regional, national and international variants, which change over time. Data structures and databases of static terms are therefore not useful to the cultural community. We need databases to reflect that meaning changes both temporally (whence etymology) and spatially, even within a culture (e.g. national, regional and local differences) and especially between cultures. For this reason traditional quests for dictionaries to provide exact equivalents in different languages have given way to new strategies that entail mappings, walkthroughs, and bridges among words and concepts. Present day semantic web models are still in terms of traditional dictionaries. Needed are models, which reflect an historical shift from traditional dictionaries (in terms of what something is) to modern versions of dictionaries that map between meanings without reducing them to a simplest common denominator. Needed is an approach that is multi-lingual and multi-cultural. [1]*

The topic of bridges between dictionaries is further discussed in chapter 8, Countermeasures.

[6] http://en.wikipedia.org/wiki/Symbol

[7] http://en.wikipedia.org/wiki/WordNet

[8] United Nations Standard Products and Services Code, http://en.wikipedia.org/wiki/UNSPSC

[9] http://www.m-w.com/

4 REFERENCE

A reference is a concept or class, its definition and attributes. In this chapter we list a set of research topics on ontology evolution, engineering and management.

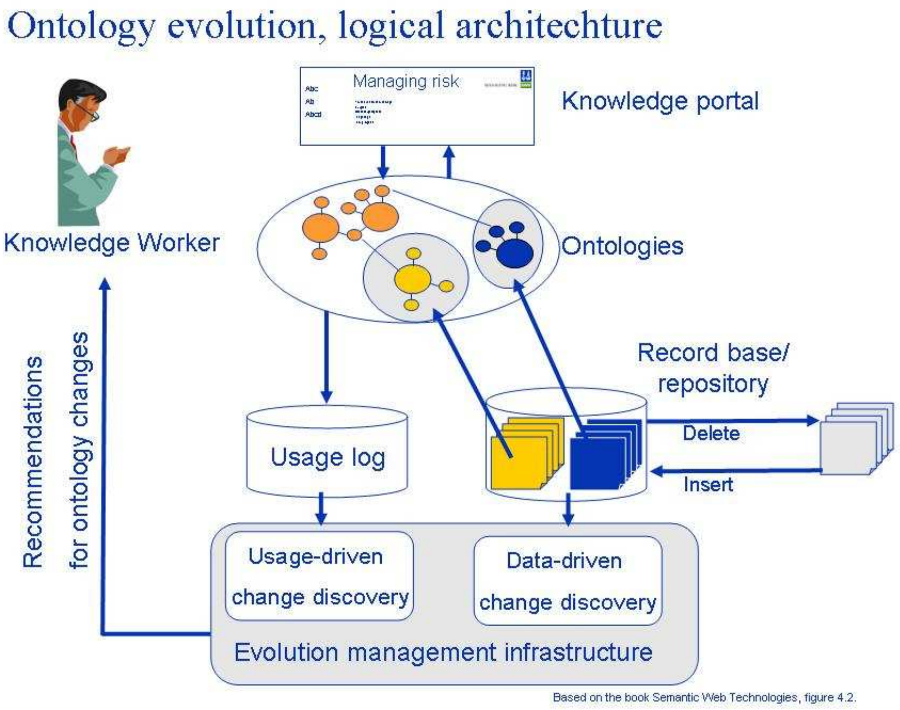
This chapter is based on the chapters “Ontology Evolution”, “Ontology Mediation, Merging and Alignment” and “Ontology Engineering methodologies” in [11] and the chapters “Theoretical Foundation of Ontologies”, “The most outstanding Ontologies” and “Methodologies and Methods for Building Ontologies” in [26].

What is ontology [26]: “*An ontology may take a variety of forms, but it will necessarily include a vocabulary of terms and some specification of their meaning. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the possible interpretations of terms”*.
Simplified, an ontology is a set of references, their definition, attributes and areas of use.

Separating the knowledge model / ontology from software and data, brings new opportunities into dynamic behaviour and process change. Technology support in these processes and the ability to do effective software development and maintenance can be improved. This ability is of great interest for both the record repositories and the software used when handling preservation of semantic value.

4.1 Ontology evolution

During its lifetime an ontology will be changed by its users. Based on inspiration of [11] this evolution process is illustrated in the figure below.


Ontology evolution, logical architechture
Managing risk
Knowledge portal
Knowledge Worker
Ontologies
Record base/ repository
Delete
Insert
Usage log
Usage-driven change discovery
Data-driven change discovery
Evolution management infrastructure
Recommendations for ontology changes

Based on the book Semantic Web Technologies, figure 4.2.

Figure 6, Ontology evolution logical architecture [11]

The two main types of changes in an ontology are *usage-driven* and *data-driven* changes.

- Usage-driven: there is a change or unbalance in how users use the archive/knowledge portal, and what support the portal can have from its ontology(ies). The use of records have changed, and the ontology needs to be changed accordingly. As input to the ontology change process, use of pattern recognition for tracking user behaviour or explicit changes in work procedures could be useful. Also search logs and analysis could be a source for capturing usage-driven changes.
- Data-driven means that the records contained in the digital repository have changed properties, volume or the new records are not good enough reflected in the ontology. Techniques from e.g. data mining and business intelligence are used to discover data-driven changes. And the results of using these techniques will be useful input to the change process.

Usage-driven and *data-driven* changes could be a type of functionality as mentioned in Figure 2, OAIS Reference Model, and its circle: “Archive, analysis tools”. The caption “record base” in the figure above is similar to the caption “digital repository” in the OAIS Reference model.

Knowledge and usage of records change. If those maintaining an ontology do not manage to reflect the changes and fail to have the ontology updated, the ontology will not serve its purpose in a semantic solution. Ontology evolution starts with identifying changes.

Based on [11, p. 52] and Figure 6 above, one way of illustrating the stages in ontology evolution is as follows:

1. Identifying changes
2. Representation of changes
3. Semantics of changes
4. Implementation of changes
5. Propagation (check if dependent ontologies, artefacts and systems are consistent after changes.)
6. Validation of changes

From Validation the circle goes back to Capturing.

4.2 Ontology mediation

Ontology mediation is used to share data between heterogeneous sources, and/or to reuse data from different knowledge bases. Established approaches for ontology mediation are ontology-mismatches, mapping, alignment and merging. As explained in the chapters below, the methods have some overlap both in what kind of problem they are meant for and the properties of outcome.

These topics are of great interest for LongRec. This is because we may be able to take a set of records (Records.T0) from time T0, the corresponding metadata and ontologies and automatically identify parts of the changes in semantics compared to record set (Records.T1) at time T1. Further these technologies could also be used in search technology for better managing search in heterogonous sources, and present search results in one relevant rating.

4.2.1 Ontology mismatches

Different types of ontology mismatches are shortly described as follows:

- Conceptualization mismatch

  - *Scope mismatch* occurs when e.g. one ontology describes patients and the other describes taxpayers. The referents they refer to overlap but are different.
  - *Granularity level mismatch*, e.g. if one ontology describes persons at most detailed level and the other ontology describes subcategories of persons.
  - *Different worldviews* of a reference. Depending on the worldview, the ontology reflecting a reference with the corresponding referents could be described differently. E.g. there may be different views of the country Palestine, its borders, its existence etc., depending on who is making and maintaining the ontology.
- Different ways of describing a reference (explication mismatch)
  - Modelling style/paradigm,
    - use of attributes versus sub-classes
    - use of point in-time versus intervals
  - Terminology mismatch
    - Identical reference in the ontologies, but the references are linked to different symbols (synonym problem).
  - Encoding mismatch occurs when one ontology uses e.g. meters for a certain reference and another uses feet or centimetres.

4.2.2 Ontology mapping

Ontology mapping takes two or more ontologies as input, and makes a new, separate ontology describing the mappings (bridge) between the sources. The source ontologies continue their separate life.

4.2.3 Ontology alignment

Ontology alignment is the process of discovering similarities between two source ontologies. Input is the relevant ontologies, and output of the process is a specification of the correspondence between the ontologies.

4.2.4 Ontology merging

Ontology merging takes two or more ontologies as input, and makes a new ontology which is the union of the sources, plus the needed links between the sources. Some of the methods for ontology merging keep the original ontologies as is, and add a bridge ontology. The source ontologies and the bridge ontology are used together as a whole in further operations like record translation or querying.

4.3 Ontology engineering

This section is based on the chapter “Ontology Engineering methodologies” in [11] and the chapters “Theoretical Foundation of Ontologies”, “The most outstanding Ontologies” and “Methodologies and Methods for Building Ontologies” in [26].

Ontologies aim to capture consensual knowledge in a generic way, and are meant to be reused and shared across software applications and by groups of people. They are usually built cooperatively by different groups of people in different locations [26]. From library science we have a tradition of systematically classifying records by thesauri, and a set of approaches has been developed [27].

In the software system development tradition, most models of data in databases have for the last 20 years been made using entity relationship models. Object-oriented programming languages have, as part of the software models, made models of the data used in the software. Ontology engineering models the knowledge or the meaning of data, and not the way data is stored, exchanged or represented.

Ontology engineering, the way we make ontologies, is a hot research topic with many variations and many different scopes and targets. We separate the domain expert having the knowledge we would like to model, from the ontology engineers who know how to capture the knowledge in ontology models. Separate tools are often used for capturing and maintaining ontologies.

An important goal of ontologies is to separate the formal model of knowledge from the records and the software used. The software tools available to average software developers still lack much of the ability to gain from this separation. This is a challenge leading to slow uptake of semantic technology.

Ontologies may give improved usability aspects related to searching, quality checking of model and data, navigation, data mining, architecture layering, reuse of data, exchange of data etc. Further we will look into different aspects of development, maintenance and support activities related to ontology engineering.

- Ontology management activities are to define/describe: Control mechanism, schedules, responsibilities, quality steps etc, maintenance and evolution procedures. These activities are similar to general activities in projects for developing systems or software of some kind.
- Ontology development activities are: Do feasibility studies, define environment/framework where the resulting ontology should be used, choose and describe methodology for how to conceptualize, formalize, represent, and implement the target ontology, describe regime for ontology population, use etc.
- Ontology support activities are: Knowledge acquisition, evaluation, system integration, ontology merging and alignment, and configuration management.

The ontology engineering methodology has similarities with the software development approach called the water fall model. Ontology evolution has similarities to iterative software development.

Further relevant research is to look into topics like:

- Framework for comparing ontology engineering methodologies, e.g. [20] “A survey on ontology creation methodologies”.
- Diligent methodology, distributed ontology engineering and frequently changing user needs.
- Ontology evaluation, e.g. OntoClean [18]
- Quality metrics for ontologies described in a separate chapter below.

4.3.1 Aspects on metamodel for Ontology engineering methodologies

Principles for how references are described are based on ontology metamodel attributes. The metamodel for an ontology gives ontology engineers different types of construction mechanisms for making an ontology. Examples of such construction mechanisms are:

- Type of formalism (how well defined are the different constructing mechanisms seen from a computational view)

- Logic
- Frames (metaclasses having other classes as instances, results are levels of classes in the ontology). (e.g. class of class constructions, could be used for categorising or grouping of classes.)
- Descriptive logic
- Types of relations
- Use of attributes on objects
- Distributed ontologies
- Fuzziness in ontologies (precision based on need and use)

A number of different methods and tools exist for building ontologies and representing ontologies. A comprehensive overview is listed in [26].

There is some discussion on strengths and weaknesses of methodologies for ontology engineering. An example of this related to OWL and description logic is presented in [19], which lists some shortcomings in ontology engineering methodology.

4.3.2 Aspects of handling space - time

> *Our space/time horizon introduces a theoretical framework where universals are not in space/time, whereas particulars are in space/time. We need to remember, however, that this framework changes with time as our knowledge of universals and particulars increases or changes. In this sense, Immanuel Kant was right: All our knowledge is in space and time. Hence, we need to go much further for two fundamental reasons. First, we need to reflect cultural and historical dimensions of knowledge. Second, we need to reflect recent developments. [1]*

Most of the existing ontologies have an upper level, above space/time horizon, and a specific level handling time and space. This two level approach asks for a very strong and consequent modelling approach. Best practice methodologies in the field are quite immature and have low uptake rate, this includes e.g. ISO 15926[39], ISO 10303 STEP [38], UML as used in the Norwegian SERES project[10].

Further literature:

- Standard Upper Ontology Working Group (SUO WG), 4D Ontology. http://suo.ieee.org/SUO/SUO-4D/index.html
- Origins of The IEEE Standard Upper Ontology. http://www.ontologyportal.org/pubs/IJCAI2001.pdf
- The Ontology of Spacetime (Philosophy and Foundations of Physics, Volume 1). This book contains selected papers from the First International Conference on the Ontology of Spacetime. http://www.spacetimesociety.org/

4.4 Ontology, identification of references

The ability to use unique identifiers when referring to a reference is a prerequisite for using ontologies in runtime environments. Further the IDs are needed in tasks of ontology alignment, merging, mapping

[10] Semantic register for electronic services, hosted at The Brønnøysund Register Center, http://www.brreg.no/samordning/semantikk/

etc. Use of OID (object identifier) or URI (Unique resource identifier) are commonly used in best practice.

4.5 Ontology quality

The article [14] “A Realism-Based Approach to the Evolution of Biomedical Ontologies” discusses ways to calculate the quality of ontology evolution. There are initiatives like OntoClean to develop methodologies for validating the ontological adequacy of taxonomic relationships. OntoClean has provided a logical basis for arguing against the most common modelling pitfalls, and argues for what we have called “clean ontologies” [18].

In the paper “Reflexive standardization: side effects and complexity in standard making” [2], Ole Hanseth addresses the general question: “What historical or contingent events and factors influence the creation of ICT standards, and in particular, their success or failure?” A standard in this context is e.g. an ontology or an EDIFACT-specification used in the health sector. The paper makes three key contributions: *(1) it demonstrates the socio-technical complexity of IS standards and standardization efforts; (2) it shows how complexity generates reflexive processes that undermine standardization aims; and (3) it suggests a theoretical interpretation of standardization complexity by using ideas from complexity theory and the theory of reflexive modernization. These research questions are addressed by offering an historical and contingent analysis of the complexity dynamics emerging from the case.*

4.6 Knowledge base evolution

In the field of knowledge management [27], the transition from personal knowledge to public and documented knowledge is a circular, continuous movement. Personal knowledge is based on public and/or some other persons’ knowledge. When new personal knowledge is documented it can become published, public and then be a part of a common knowledge base. In the Knowledge Management Field, knowledge can be recorded in records and made public available.

> *DIKW is the proposed structuring of data, information, knowledge and wisdom in an information hierarchy where each layer adds certain attributes over and above the previous one. Data is the most basic level; Information adds context; Knowledge adds how to use it; and Wisdom adds when to use it.*[11]

From a LongRec perspective this means that the documented and thereby explicit knowledge in principle is accessible and can be part of a semantic analysis. But we miss the personal knowledge of those who made and performed operations on records in the past.

4.7 Ontology distribution and distributed knowledge base

The knowledge base consists of the information stored in the records themselves and in the relevant ontologies. Additionally, referral may be made to records maintained by external parties. Thus, one may need to rely on the continued support and trustworthiness of external parties and their record management regimes. This situation is relevant to both distributed ontologies and record archives which are not self contained. Linkage between records located in different locations will be challenging to maintain.

[11] http://en.wikipedia.org/wiki/DIKW

4.8 Best practice

In a LongRec case study at the Brønnøysund Register Centre uses entity relationship models for their system implementation. However, recently an UML model of the references has been made, based on a temporary UML Profile for ontology engineering.

The Brønnøysund Register Centre has an UML Profile describing how UML should be used to describe the ontology in their SERES, “SEmantic Register for Electronic Services”. Further the Norwegian Centre for Informatics in Health and Social Care (KITH) has its own UML profile.

UML is commonly used as best practice, and it is therefore interesting to note that [26, chapter 1.3.3] discusses use of UML methodology, notation and tool support for ontology engineering.

KITH uses OIDs for identifying concepts. E.g. ID-types of organizational categories (OID=9051). Brønnøysund Register Centre uses both unique paths and numbered IDs for classes in the UML. Code lists have a numbered ID. Both classes and code lists have version numbers.

5 REFERENT

A referent is an instance of a reference and it has a lifespan.

5.1 Life cycle standards

This chapter lists some standards used for keeping engineering and life cycle support information about a referent in the different stages of its life cycle, e.g. information about a ship during its lifetime.

Examples of standards used for modelling information about referents during a life cycle are PLCS [38] and ISO 15926 [39]. The history of the standards goes 15-20 years back. The standards are in some use in special domains.

Examples of current research in the field are reflected in [16] “Strategies for Referent Tracking in Electronic Health Records”. The article elaborates the challenges of keeping all patient information in such a formal shape that it is suitable for search, statistic production of statistics and reasoning.

5.2 ID of referents

Semantic Technologies like Topic Map [40] and RDF [41] (Resource Description Framework) use URIs (Uniform Resource Identifiers) to manage the difficult task of identifying both references and referents in a distributed environment.

Seen from the referent corner of the triangle an illustrated example of the problem is how to make sure that a record about a referent is unambiguously identifying the correct referent.

In a record many symbols are used, and many pointers to references are made. So if the main purpose of a record is to describe something relevant for a patient, many other references are made. Let us look into an example in an electronic patient record system

1. Record 1: The patient has a tumour type AAA in his left kidney, at time t0
2. Record 2: Kidney removed by surgery of type x, at time t1, by Dr. Inger Mette Gustavsen. Medicine of type A and B shall be used.
3. Record 3: Kidney transplantation at time t2. Medicine of type A and B shall be used.
4. Record 4: Transplantation failed after two months.
5. Record 5: Kidney transplantation, t3. Medicine of type A and C shall be used.

Among others the symbols “tumour”, “kidney” and the identification “left” are used. For each record there will be a linkage between the record and the patient.

All the records are linked to the patient but not properly linked to the doctor or type of medicine used and only in one of the records “left kidney” is used. Are we sure that all the records are related to left kidney? Most records have a timestamp, but also periods are used like “after two months”. Further details in the use of medicines are listed in other documentation, and a set of answers from laboratories are separate records or attached to a record.

Suggestions for how to improve the situation is made in articles like [16] “Strategies for Referent Tracking in Electronic Health Records”.

5.3 Best practice

Brønnøysund Register Center uses global unique identifiers for companies as referents (Norwegian legal companies). The ID is called "organisasjonsnummer". But the challenges remains when other companies identify Norwegian companies by e.g. name alone. Another challenge is that a legal company may be located at several sites. An ID for each site or activity is established and is called "bedriftsnummer".

It is often difficult to make precise links between symbols and referents. There is a many-to-many relationship between symbols and references. E.g. different languages have a word for the reference "tree", and in one language there may be synonyms for the symbol "tree". Similarly one referent may be categorized by different references depending on worldview. Since many referents may be categorized in the same category, there is also a "many to many" relationship between reference and referent. This means that the path from symbol – reference – referent is not unique for identifying one specific referent, and we need a separate identification regime for referents.

The identification challenge of referents makes it difficult to design services based on content of records, since most records consist of symbols, not identifiers to referents. This is one important reason for why it is challenging to establish good search rating and functionality, business intelligence, statistics and reasoning on records related to referents. Improvements in how the content of the corners in the triangle of Ogden [3] are identified and especially how referents are identified, will improve data quality and the ability to make high quality semantic services on records.

6 RECORD

A record is text, pictures, numbers etc. describing some aspects of a referent(s). In this chapter we present some issues related to a record itself and how it is influenced by time.

6.1 Data quality

Research within data quality and completeness of information operates with e.g. levels of what records conform to, and lists a set of criteria to be fulfilled in order to become conform. ISO 2382-8 defines data quality as follows:

> *Data quality is "the correctness, timeliness, accuracy, completeness, relevance, and accessibility that make data appropriate for their use"*

Different levels of data quality are categorized in [23] as:

> *"the three semiotic levels—syntactic, semantic, and pragmatic—describing respectively (1) form, (2) meaning, and (3) application (i.e. use or interpretation) of a sign can be used to define corresponding quality categories based respectively on (1) conformance to database rules, (2) correspondence to external (e.g. real-world) phenomena, and (3) suitability for use."*

Compared to the terminology in this report:
(1) Conformance to database rules is relevant in relation to symbols, and if the database rules have semantic impact (e.g. not only structures to avoid redundancy or database administrative rules).
(2) Correspondence to external (e.g. real-world) phenomena are linked to both reference and referent
(3) Suitability for use is taking the challenge a bit further than the scope of this report. In this report we focus on understanding the content of records, but may end up in situations where we understand the content, but the content of records are not complete enough to actual perform the tasks we would like.

In [24] "Journey to Data Quality" methodologies and quality metrics based on research are presented. This is presented along with case studies and suggestions on how to close the gap between the current quality level and the level you wish. An interesting overview of "Ten root conditions of data Quality" is elaborated, and consequences are discussed. The authors suggest treating information about a product as a product itself, not a bi-product, and present methodology and notations to make information product maps. The technique is used on example cases.

A relevant standard initiative is IS0 8000 Data Quality [44].

A reflection: Every record is a referent, and also every ontology is a referent.

6.2 Data and model transformation issues

Once material is digital, it can be translated to other formats, representations and media. This is called transformation of a record. But also models and ontologies can be transformed from being built according to rules of one metamodel and transformed to be according to another metamodel [see e.g. OMG MOF [42]). This holds true for many cases, but e.g. semantics contained in a written speech in a play, will not capture the speakers way of performing the speech. A video of a play will contain more context information than an audio recording of the same play, but the video does not contain the theatre atmosphere and building architecture, audience, smell etc. Much of the semantic is in the way a speech is performed, the context the speech is held in, gestures of the speaker etc. This means that a digital text record of a speech contains less or at least other semantics than an audio file of the same speech, and also less semantics than a video of the play.

On the other hand, the audio part of a video could be used as a complete audio file.

A digital report can become a printed paper copy, with (under some circumstances) identical attributes to other paper copies.

Based on the chapter above we believe transforming could change the semantics of a record. Since many preservation regimes have transformation as one of several mechanisms for managing records, the semantic preservation should not be neglected during transformation.

6.3 Metadata

The article [15] "Introduction to Metadata, Pathways to Digital Information" describes from a metadata world point of view, the different aspects of the life and use of metadata. Types of metadata listed in [15] are as follows:

Type	**Definition**	**Examples**
Administrative	Metadata used in managing and administering information resources	- Acquisition information - Rights and reproduction tracking - Documentation of legal access requirements - Location information - Selection criteria for digitization - Version control and differentiation between similar information objects - Audit trails created by recordkeeping systems
Descriptive	Metadata used to describe or identify information resources	- Cataloguing records - Finding aids - Specialized indexes - Hyperlinked relationships between resources - Annotations by users - Metadata for recordkeeping systems generated by records creators
Preservation	Metadata related to the preservation management of information resources	- Documentation of physical condition of resources - Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration
Technical	Metadata related to how a system functions or metadata behave	- Hardware and software documentation - Digitization information, e.g., formats, compression ratios, scaling routines - Tracking of system response times - Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of information resources	- Exhibit records - Use and user tracking - Content re-use and multi-versioning information

Supplementary types could be metadata for presentation, references to ontologies, geographical context, information governance regime, purpose, intention of writer, intellectual property rights etc.

6.4 Best practice

The OAIS, Reference Model for an Open Archival Information System [35] is regarded as a best practice system for archival systems. The LongRec partner The National Library of Norway uses the OAIS as basis for their preservation architecture.

Other relevant standards for archival systems and practice are Noark [30] and Moreq [31].

Relevant standards for best practice for metadata are

- ISO 11179 Information technology Metadata Registries [34]
- Dublin core [33]
- DDI, Data Documentation Initiative [32]. DDI is an effort to establish an international XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioural sciences
- Semantic annotation WSDL by W3C [45]

7 WORLDVIEW AND CONTEXT

7.1 Worldview

> *The semantic web communities focus on what a thing is (its substance), and do not allow for a gradual historical shift from substance to function. As a result the AI and semantic web communities create data structures that assume a single world-view. Every thing is presented as if this is the way "it is" ontologically, rather than providing frameworks whereby what a thing "is", what it means, and how it relates to other things, change as the framework changes. This dimension is needed a) to explore the interplay between facts and the frameworks or world-views used to explain them and b) to explain a historical shift from a quest for a single ontology to a need for multiple ontologies. Needed is an approach where entities can evolve in meaning. [1]*

The triangle of reference [3] has evolved, and different alternative views of the triangle have been made [1, p 17]. However, even if we define our concepts/references separated from the terms/symbols and referents, we are influenced by our worldview and our purpose of modelling.

If a border of a country is disputed, then there will be different worldviews of the border. The same is the case in a legal dispute, e.g. if a certain delivery meets the completion criteria in a contract or not. The two parties may have different views of the concept of what should be delivered according to an agreement. This leads us to the need for managing multiple concepts in ontologies, and to the need for a mechanism for managing ontology merging and comparing.

The author of a record had an intention writing it. The readers knowledge of the writers' intention will influence which worldview he will try to use as basis for his understanding the content.

Another example could be found in the feminine literary history. The knowledge base before feminine literature became available was heavily influenced by men's view of the world. We now lack the feminine worldview in a large part of our cultural history; we more or less have the history and worldview of men [28].

A systematic approach for managing ontologies with different worldviews covering the same referents would be of great interest for LongRec. The methods found so far on ontology merge and mapping may be used for handling several ontologies covering the same referents, but with different worldviews.

7.2 Context information

> *Since every culture focuses on some aspects of knowledge and ignores others, history is essential to understand both the sources of our views and the limitations of the frameworks or worldviews with which we present them.* [1, p9]

Handling context of a referent is related to whether the referent is material or not. *A material object* has a physical life-cycle, and information related to the referent will usually relate to the referent in a certain state or usage in a certain phase of its lifecycle. Examples of material referents could be a ship, a book or a person. A material object can not be fully digitally represented, but information about the object may be digitalized. A digital copy may be transformed to a material object, e.g. the digital original of a book can be printed and become a physical copy of the book. The whole content of a book could be digitally represented, but the physical book will still be physical and have some properties different from the digital copy.

A immaterial referent could be an house insurance, email, bank transaction or a video stream. The house is material, the insurance contract can be transformed to a material paper record, but the house insurance in it self is not physical[12]. An email can be printed, and than the printed representation of the email is physical. Characteristics of the *immaterial* objects are that they do not have any physical representations in real life that the record can relate to. To the LongRec project this means that we will focus on a total conservation regime comprising objects, metadata, ontologies, context etc.

7.3 Semantic dimensions

Inspired from the field of philosophy the six basic questions – Why? How? When? Where? Who? What? – could have been used to structure this report. Some literature does focus on this problem breakdown structure.

> *Libraries typically offer access via author catalogues (Who?) and title catalogues (What?). In addition, libraries such as the Herzog August Bibliothek in Wolfenbüttel, offer access chronologically (When?) and via locations of publication (Where?). Search engines such as Artefacts Canada have begun to use such questions for searching. [1]*

Systematic search using the six basic questions and their variants would greatly expand the scope and the precision of searching and could be of great help as countermeasure for preserving semantic value.

[12] This report does not go in depth in discuss what material is, or what it is not. We believe the answer depend on worldview.

8 COUNTERMEASURES TO LOSS OF SEMANTIC VALUE

Countermeasures are used to slow down the deterioration of semantics and/or log the semantic changes for later being able to inform coming users of records. This chapter is split into (i) Procedural and administrative countermeasures and (ii) Technology based countermeasures. The examples listed may contain parts from both i and ii.

8.1 Procedural and administrative countermeasures

8.1.1 Distributed knowledge base

Linkages between records maintained by different parties identify the challenge of managing distributed management regimes for record preservation. At least the three corners symbol, reference, referent and the record themselves will be part of this challenge. To some extent, this challenge is handled in archival standard initiatives like Noark and Moreq.

The LongRec case studies may identify a need for further elaboration on these topics.

8.1.2 Conservation regime for semantic value

A challenge will be to find the semantics in old records. Some of our challenges in LongRec are related to records not held by a proper conservation regime. This means that we have no or little explicit information about the context or worldview of the author, and little or no metadata.

Based on our literature search it seems unclear if the value of the information from knowledge engineering in the 1980s and 1990s has been meaningfully preserved. This could be an example of the digital bomb, i.e. records relevant for this report, which we were not able to find, read and understand.

8.1.3 PANIC: A conservation regime

Semi-Automated Preservation and Archival of Scientific Data using Semantic Grid Services is described in [5], and is called PANIC. This is an advanced service offered for long term preservation and parts of it is using semantic technology.

> *Moreover, it is generally recognized that there is no single best solution to digital preservation. Differences in the needs and practices of various scientific disciplines make it difficult if not impossible to define a 'one size fits all' approach to selecting, appraising and retaining scientific data. The most appropriate strategy depends on the particular requirements of the custodial organization, the producers and users of its collection and the nature of the objects in the collection. Hence within the PANIC project we combine the efforts of the different domain-specific preservation initiatives by integrating the range of tools and services being developed into a single encompassing Grid framework. More specifically PANIC uses a flexible, dynamic, semi-automated approach which provides access to a range of metadata tools and risk assessment, notification, emulation and migration services through a Semantic Web/Grid services architecture.* [5]

The PANIC system comprises three main components:

- Preservation metadata generation tools
- Obsolescence Detection and Notification services

- Preservation Service Description [5]

8.2 Technology based countermeasures

8.2.1 Time series of ontologies for preservation of semantics

Records need several types of metadata, including definition/ontology metadata, to be useful. In the LongRec focus of preservation of semantic value, the result of our literature study mainly shows us a snap-shot of the last version of an ontology. Further literature study will hopefully show some initiatives on this topic, candidates could be:

- Eurostat's Metadata Server http://ec.europa.eu/eurostat/ramon/
- http://www3.ssb.no/stabas/DOCS/Neuchatelversion2.1.pdf
- The DDI initiative [32].

8.2.2 Automatic building of ontologies

An important research topic is ontology learning that aims at providing (semi-) automatic support for building ontologies. This aspect is addressed in [11, chap 2] and [17, Chap 9].

8.2.3 Semantic annotation

The process of linking ontologies and record content is referred to as semantic annotation. This linkage is basis for value adding services and advanced search. By adding metadata about semantics to records, software functions and search engines may use both the record content and the semantic annotations to give the users a more sophisticated level of services.

Annotations could be:

- Manually added by a record owner or communities of users or other interested parties
- Added by algorithms either based on record analysis, ontology/reasoning analysis, user behaviour analysis or other changes.
- Combinations where e.g. algorithms suggest annotations and users approve, edit or redraw them.

APPENDIX A: TERMS AND ABBREVIATIONS

Semantics	The science of meaning
Semantic (adjective)	Of or relating to meaning in language. (Webster)
Semantic technology	Semantic technologies consist of • A model for explicitly encoding the meaning of information • Software for using and maintaining the model. These models are often called ontologies.
Semantic solution	Semantic solution consists of • Semantic technology • Records related to the model Software services with the purpose of offering functionality to system or users where the use of models are an important part.
Ontology	An ontology may take a variety of forms, but it will necessarily include a vocabulary of terms, and some specification of their meanings. This includes definitions and an indication of how concepts are inter-related, which collectively impose a structure on the domain, and constrain the possible interpretations of terms [26]. An ontology is typically a set of references.
Semantic model	Within the term Semantic Model we include the • Designed ontology, or the implied folksonomy • The instantiated level • The semantic annotations, i.e. linkage between the ontology and the instantiated level.
Semantic value	Is the amount of precision, quality and completeness in a record, in its corresponding types of metadata and in related records. With the purpose of making users able to interpret the content correctly to their purpose. Ideally semantic value should be possible to measure, and a high semantic value should make it easy for a user of a record to interpret it according to writer's intention, worldview, reference, symbols etc. Metadata should be related to all the ten areas of change listed in this report.
Semantic annotation	Semantic annotation is a relation between a record, or content in a record, and an ontology. The annotations are explicit and have metadata related to author, creation time etc.
Record	It is a container keeping data and knowledge about a referent. Could be structured or unstructured text, pictures, numbers, model etc. describing some aspects of a referent. The record should have a clear boundary and contains symbols. It may have the following properties: content, structure (reference to an external model e.g. an XSD), metadata (e.g. Dublin Core), presentational metadata and a technical format for storage and exchange.
OAIS: information packages	*The Content Information and associated Preservation Description Information which is needed to aid in the preservation of the Content Information. The Information Package has associated Packaging*

	Information used to delimit and identify the Content Information and Preservation Description Information.
Knowledge base	The sum of records and ontologies represented in such a way that it is useful to people or machines.
Symbol	The words in a language or symbols like traffic signs. Symbols are used to identify a reference.
Reference	A concept or class, its definition and attributes.
Referent	An instance/object, which has a lifespan.
OAIS: Content Information:	*The set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc.*
OAIS:Information Object	*A Data Object together with its Representation Information.*
OAIS: Data Object	*Either a Physical Object or a Digital Object.*
Data quality	ISO 2382-8 defines data quality as follows: Data quality is *"the correctness, timeliness, accuracy, completeness, relevance, and accessibility that make data appropriate for their use"*.

APPENDIX B: REFERENCES

1 Towards a Semantic Web for Culture, 2004, Veltman. http://jodi.tamu.edu/Articles/v04/i04/Veltman/veltman.pdf

2 Reflexive standardization: side effects and complexity in standard making. Ole Hanseth. http://heim.ifi.uio.no/~oleha/Publications/misqsi3979r2.pdf

3 The Meaning of Meaning, C. K. Ogden & I. A. Richards, sixth edition. 1943.

4 Approaches for Semantic Interoperability between Domain Ontologies. B. Orgun, M. Dras, A. Nayak. ACM.

5 Semi-Automated Preservation and Archival of Scientific Data using Semantic Grid Services. Jane Hunter and Sharmin Choudhury. Semantic Infrastructure for Grid Computing Applications Workshop at the International Symposium on Cluster Computing and the Grid, CCGrid 2005. Cardiff, UK. May 2005

6 PADI is a subject gateway to international digital preservation resources. http://www.nla.gov.au/padi/index.html

7 TRUSTWORTHY 100 – YEAR DIGITAL OBJECTS. Syntax and semantics - tension between facts and values. H.M. Gladney. (is an unpublished draft provided for critical discussion and no other purpose.)

8 SAWSDL Semantic Annotations for WSDL. W3C recommendation 28. august 2007.

9 OWL 1.1, http://www.w3.org/Submission/owl11-overview/

10 Challenges for the Semantic Web and Information Systems from Culture, 2006, Kim H. Veltman. http://www.nla.gov.au/padi/topics/32.html

11 Semantic Web Technologies. Trends and research in ontology-based systems. Editors: John Davis, Rudi Studer, Paul Warren. Wiley 2006.

12 ISO 1087, Terminology work, Vocabulary. Part 1: Theory and application

13 A Literature Review for the Problem of Biological Data Versioning. Ben Tagger Started - 28th July 2005. http://www.cs.ucl.ac.uk/staff/btagger/LitReview.pdf

14 A Realism-Based Approach to the Evolution of Biomedical Ontologies. Werner CEUSTERS , Barry SMITH. forthcoming in Proc. AMIA Symp. 2006. http://ontology.buffalo.edu/bfo/Versioning.pdf

15 A.J. Gilliland-Swetland. Setting the Stage. Introduction to Metadata, Pathways to Digital Information. 2000. http://www.getty.edu/research/institute/standards/intrometadata/

16 Strategies for Referent Tracking in Electronic Health Records.
Werner Ceusters, Barry Smith. Preprint version of paper forthcoming in Journal of Biomedical Informatics

17 Handbook on Ontologies. Springer Series on Handbooks in Information Systems S. Staab, R. Studer. 2004.

18 An Overview of OntoClean by Nicola Guarino and Christopher A. Welty. http://ontolog.cim3.net/file/resource/presentation/OntoClean--ChrisWelty_20041118/guarinowelty_final_v4.pdf

19 Ontology and Medical Terminology: Why Description Logics Are Not Enough. Werner Ceusters, Barry Smith, Jim Flanagan. Towards an Electronic Patient Record (TEPR 2003),

20 A survey on ontology creation methodologies, Cristani, M; Cuel, R. 2005. International Journal on semantic web and information systems.

21 Towards A Realism-Based Metric for Quality Assurance in Ontology Matching. Werner CEUSTERS. Center of Excellence in Bioinformatics and Life Sciences, Ontology Research Group

22 SERES, SEmantic Register for Electronic Services. http://www.brreg.no/samordning/semantikk/samarbeid.html

23 A Semiotic Information Quality Framework, Rosanne J. Price, Graeme Shanks. The IFIP TC8/WG8.3 International Conference 2004

24 Journey to Data quality, Yang W. Lee et al. MIT Press 2006.

25 A Comparative Study of Semantic Technologies, version 1.0 Norstella 2007.

26 Ontological engineering, Asuncion Gomez-Perez et al, Springer 2003.

27 Using and preserving corporate knowledge during times of change. Johanna Gunnlaugsdottir, Procedings of the 12th Nordic conference for information and documentation, Knowledge and Change. 2004.

28 XML and Document lifecycle. Focus on digital preservation, Master Thesis by K. Helen Møller. Oslo university college, faculty of journalism, library and information science.

29 Ways of reading. Martin Montgomery, et al. 3rd edition 2007. Published by Routldge.

30 Noark, a specification of functional requirements for electronic recordkeeping systems used in public administration in Norway. http://www.arkivverket.no/english/electronic.html

31 Moreq, Documentation on Model for Electronic Record Management

32 DDI, Data Documentation Initiative. http://www.icpsr.umich.edu/DDI/

33 Dublin Core, http://dublincore.org/

34 ISO 11179, Information technology Metadata Registries

35 OAIS, Reference Model for an Open Archival Information System. Januar 2002.

36 Leveraging Semantic Technologies for Enterprise Search, by Gianluca Demartini. Proceedings of the ACM first Ph.D. workshop in CIKM. 2007.

37 Enterprise Search Report 2008, Comprehensive Product Evaluations. CMS WATCH.

38 PLCS ISO 10303 “Product Life Cycle Support

39 ISO 15926 “Industrial automation systems and integration—Integration of life-cycle data for process plants including oil and gas production facilities”.

40 Topic Maps ISO/IEC 13250:2000 "Topic Maps" standard

41 RDF http://www.w3.org/RDF/

42 OMG MOF http://www.omg.org/mof/

43 UML http://www.uml.org/

44 ISO 8000 Data Quality developed by ISO TC184/SC4

45 SAWSDL http://www.w3.org/2002/ws/sawsdl/

APPENDIX C: LITERATURE SEARCH

Based on input from project participants we performed a literature search the 15th of August 2007 in the Inspect database using the terms listed below. Manually we read through the abstracts of #4, #5, #9 and #19. In addition a number of papers, books and reports and their reference lists have been basis for our literature search. Some results have also come out of internet search engines.

#19 (state of the art) and ((("knowledge-representation" in DE) or ("semantic-Web" in DE)) and (PY > 2001))(39 records)
#18 (("records-management" in DE) and (PY > 2001)) and (state of the art)(4 records)
#17 state of the art(24167 records)
#16 (("records-management" in DE) and (PY > 2001)) and (state of the art and (TR:INSP = REVIEW))(2 records)
#15 state of the art and (TR:INSP = REVIEW)(6900 records)
#14 (("records-management" in DE) and (PY > 2001)) and (TR:INSP = REVIEW)(337 records)
#13 (review and (PY > 2001)) and ("records-management" in DE) and (PY > 2001) and (TR:INSP = REVIEW)(25 records)
#12 (review and (PY > 2001)) and ("records-management" in DE) and (PY > a 2001)(77 records)
#11 review and (PY > 2001)(139711 records)
#10 ("records-management" in DE) and (PY > 2001)(1378 records)
#9 ((old data) in AB) and (PY > 2001)(81 records)
#8 ((old data integration) in AB) and (PY > 2001)(0 records)
#7 ((krogstie) in AU) and (PY > 2001)(12 records)
#6 (("knowledge-representation" in DE) or ("semantic-Web" in DE)) and (PY > 2001)(5855 records)
#5 (Records-Management-Bulletin in SO) and (PY > 2001)(179 records)
#4 (records preservation) and (PY > 2001)(12 records)
#3 (record semantics) in AB(2 records)
#2 "data-integrity" in DE(6398 records)
#1 "records-management" in DE(4472 records)