LongRec

Long-term Records Management

Research Results from Project Period 2006-2010

LongRec
0100111 0100110I

MANAGING RISK
DNV

A typical data center currently spends

US $75,000

annually to power and cool a high-end system supporting 1 Petabyte of data

(@ IEEE in Computing, 12. nov 2008)

More than 50.000 file formats

On a macro-level, obsolescence of file formats, hardware, and software will be the dominant challenge of the future. The current explosion in the amount and heterogeneity, i.e. multimedia, documents, programs etc., of data generated annually will put search and retrieval technologies to the test.

On a micro-level, records management will face risks associated with long-term issues, such as data protection over decades, loss of reputation, trustworthy archiving, etc.

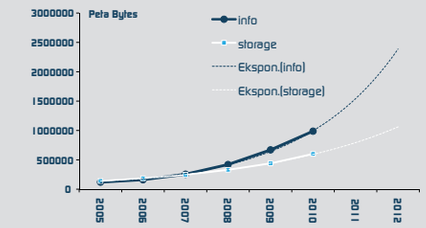
File format obsolescence

Adobe · Microsoft Office · WordPerfect · Ps

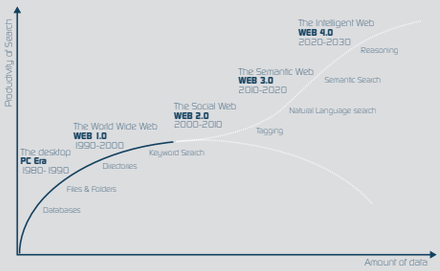
Propietary, closed specifications, e.g. Word.doc
Evolve quickly, exist in many different versions for different platforms, with only limited backward compatibility

Propietary, open specifications, e.g. Adobe.pdf
Vulnerable to market forces as they can be abandonded for commercial reasons

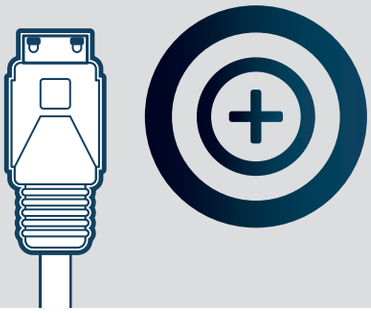
Non-propietary, open specifications, e.g. JPG
Guaranteed long-term availability, specifications published by international standard bodies. BUT these standards must be widely adopted by both user and developer.



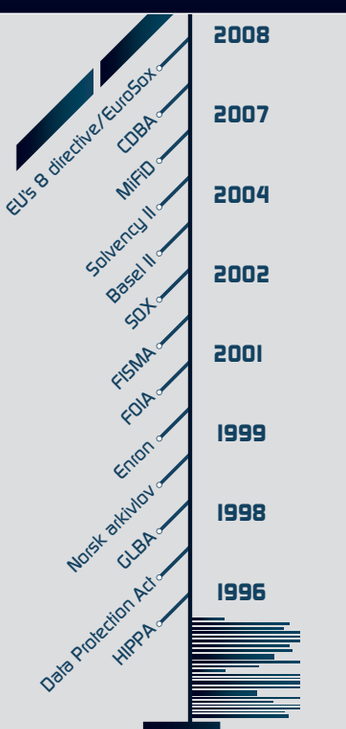
The annually expected growth in global data generation and storage capacity shows clearly that not all data generated can be stored. If this trend continues then by 2012 there will be storage space only for half of the newly generated data.



More digital signing



Increased Number of Laws and Regulations

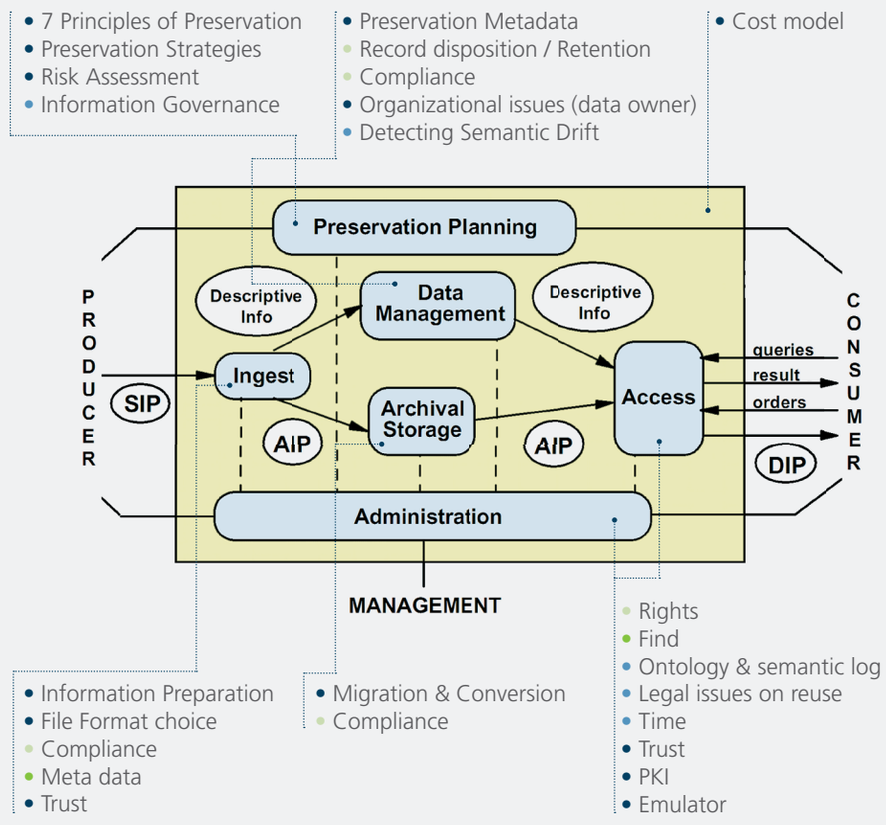


The information outlives the information carrier!

Table of content

LongRec in Relation to OAIS

LongRec is the joint-industry project focusing on the challenge of persistent, reliable, and trustworthy long-term storage of digital records, with emphasis on availability and use of information. Problems associated with these aspects typically emerge when document lifetime exceeds 20 years. This 3-year R&D project was supported by the Norwegian Research Council, financing 3 PhD students and one MSc student. LongRec had 5 major focus areas: READ - readability over time, including migration and conversion, FIND -search and retrieval, TRUST – trustworthiness and evidential value, UNDERSTAND – semantics, and COMPLIANCE. Additionally, Cost factors of the information management were addressed as a separate activity. (Ref: 1, 14, 41-58)



READ

LongRec activities in this area fitted well with the existing OAIS framework. They where mainly related to the functional entities Ingest, Preservation Planning, Archival Storage and Data Management. LongRec emphasis was on Preservation Strategies (8, 9, 21, 25, 26), Meta Data (11), Migration and Conversion (6, 7, 8, 21, 24).

COMPLIANCE

Compliance with all relevant laws and regulations require secure tracking and storage of relevant metadata during ingest as well as at the deletion of the information from all relevant media (archival storage). The main part of the OAIS framework relevant to compliance is the Data Management part. Different relevant policies have to be followed by each company – all related to compliance issues (14, 26)

FIND

Finding digital objects is not only a matter of searching but it depends to a very great deal on the availability of good indexing data. Good structured, machine-readable meta and content data will not only increase the search performance but also the understanding of the information.

LongRec focused on semantic technologies and their application as well as the temporal aspects in information search. These Find activities were mainly related to the Ingest and Access entities in OAIS. Ref (2, 3, 22, 30, 31, 34).

UNDERSTAND

The meaning of information conveyed in digital records changes over time. Moreover, the reference data pointed to by the record may also change, become invalid or disappear. LongRec investigated how Semantic Technologies can contribute to the preservation of the semantic value of information for better understanding by future users as well as better preparation by the information professionals.

Understand activities are mainly related to the Administration in the OAIS (4, 5, 12, 18, 19, 22).

TRUST

The preservation of trust over decades involves not only the records as such but also the measurement of the trustworthiness of the digital repository and the associated processes. A way to measure the trustworthiness of a record is by assessing its potential value as an evidence. This includes the digital objects, its origins, functions, actions performed, their content, form, creation context, laws governing these record types, file format, as well as other kinds of relevant metadata.

Trust activities relate to all OAIS components (10, 15- 17, 21, 23).

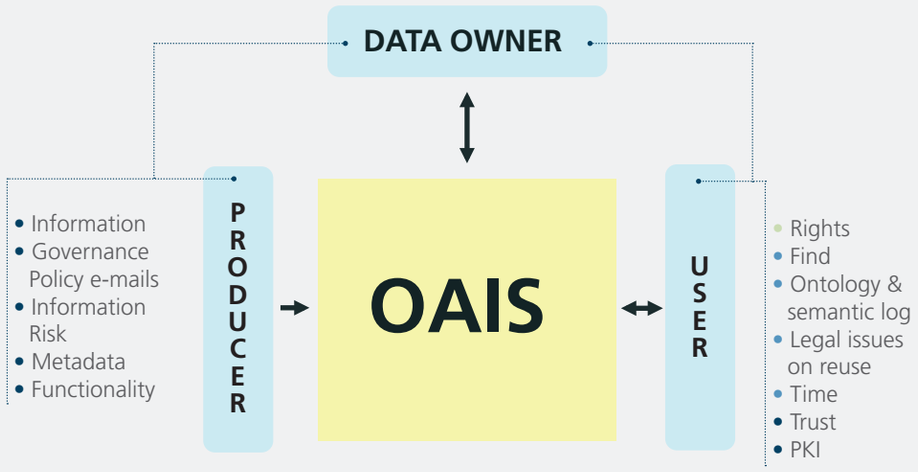
LongRec goes beyond OAIS

LongRec aimed to go beyond the digital preservation areas addressed by the OAIS. Information is often used (retrieved, updated, and verified) over a long time period. For instance, physical objects such as ships, drilling rigs or buildings have expected life times of many decades and undergo modifications with irregular frequency.

This will have to be reflected in the updates and adjustments related to ownership, authorizations and other properties. The effect of time leading to changes in the meaning of terms will make it particularly difficult not only to phrase a successful query but also to interpret the results correctly. Data users will increasingly demand services where ontologies and semantic technologies will play a central role. (Ref. 14, 43-51)

READ

Data preservation starts already at the information generation stage. LongRec went beyond OAIS as it shed light on organizational variables (of the data producer) that may have a constraining effect on preservation parameters or strategy such as retrieval frequency, retention time or separation of data from functionality. (1, 14, 16, 32)



COMPLIANCE

This is a broad domain linking several knowledge areas within a company. The OAIS framework captures compliance only partly. To follow up compliance issues related to information management one would benefit from having a virtual organization in place which would contain and track the knowledge in laws/regulations and the impact these have on business. The information management team and the IT team will have to put such demands into operation (Ref: 8, 13, 14).

FIND

From a long-term perspective, time becomes one of the most important characteristics of a data record.

The indexing of information will be increasingly difficult, not only due to the characteristics of non-textual media types but also because search indices sooner or later will have to explicitly take time into account – but what does time mean anyway. Temporal taxonomies may become increasingly important (Ref: 3, 27-31, 34)

UNDERSTAND

To understand an information object in a repository both human and machines needs supplementary information. This makes up a network of semantic links between repositories under the control of diverse information governance regimes using different technologies.

The Understand pilot used Norwegian national master data from the enterprise register and other public sources to test mechanisms for managing and visualizing distributed information repositories.

Some of the principles from the Semantic Web and Linked Open Data was used. (5, 12, 20,

TRUST

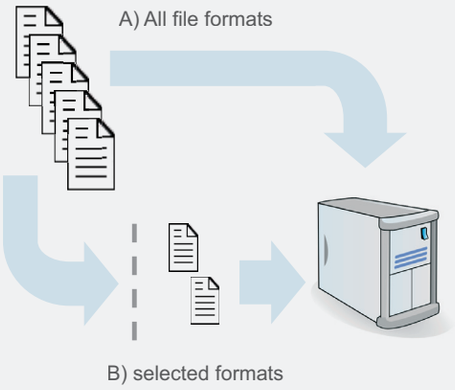
To be able to trust the digital material to be archived, enough evidence have to be available at the time of ingest (entrance into the OAIS archival environment). The collection and maintenance of evidence have to be an integral part of records management prior to digital archiving. Otherwise some of the evidence will be lost before the time of archiving. The Trust pilot (case study) has been looking into creation and maintenance of trust in the records management environment, within the ingest framework defined by the Noark standard. Various trust strategies have also been investigated covering diverse approaches for making digital material trustworthy. These strategies can be used both during records management and digital archiving (Ref: 1, 21).

TORMENT OF FORMAT CHOICE

The number of registered file extensions exceeds 18.400 – most of them proprietary formats. In a long-term perspective almost all will become obsolete requiring format conversion or other strategies to assure readability.

Two file format strategies represent the extremity: A) allow all file formats, or B) allow only a small set. A) requires a bit preservation strategy combined with keeping necessary hardware and software, or applying emulators. In B) obsolete files are converted into suitable new formats. In a combination of A) and B) only a small subset of all the formats will be converted. LongRec prepared a state-of-the-art report giving an introduction into storage media, long-term file formats, and conversion and migration issues, together with a list of recommended practices and strategies.

The National Archives and BBS apply strategy A), whereas the National Library of Norway has to use the combined A) + B) strategy facing the drawbacks from both approaches (Ref: 6, 8, 14, 57).



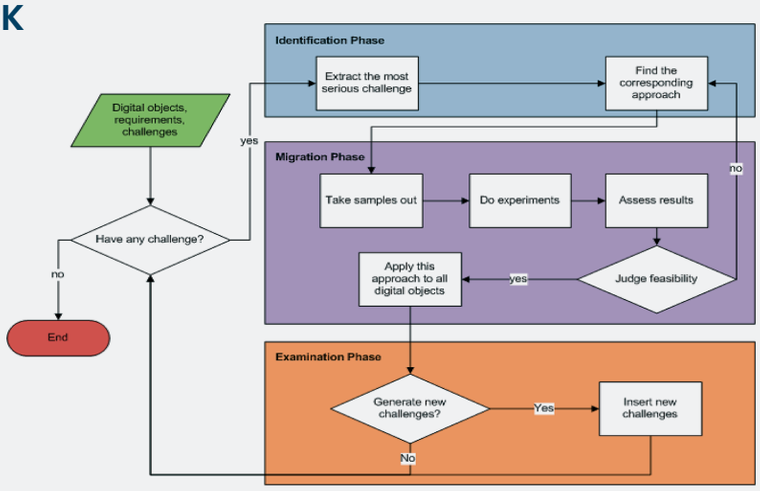
The preservation strategy depends on the chosen long-term storage file format

MIGRATION FRAMEWORK

Digital information should survive several migrations over time therefore a complete and accurate migration framework would be advantageous.

Based on an analysis of several international project a possible migration framework can be derived that is divided into the identification, migration phase and evaluation phase. Identification phase: where all potential challenges are ranked regarding their severity.
A severity score could be defined as:
Severity score =
affected scope * recoverability

Based on this score the preferred approach can be selected.

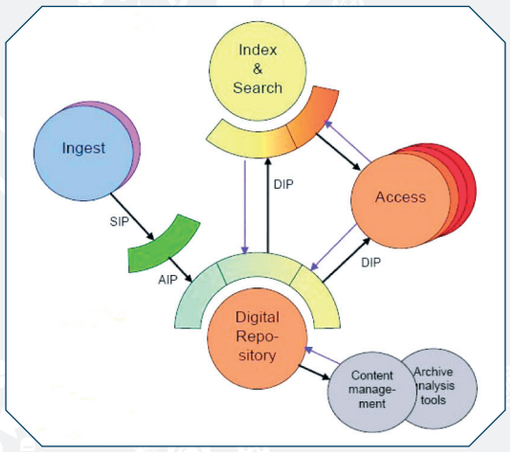


MIGRATING HUGE VOLUMES OF HETEROGENEOUS DATA

This case study addresses the challenges faced by the National Library of Norway regarding content management, primarily migration and conversion, of the National Library's trusted digital repository. The main question is how to be certain that the repository records remain unaltered, i.e. 'survive', despite multiple transitions between hardware and software which are inevitable in the future.

The complicating factor for this case is the huge volume and heterogeneity of digital content the National Library operates with, which includes not only written material but also images, various audio files, films and Internet publications (.no domain). Due to the volatility of storage media and technology and due to the novice technology that appears all information objects are migrated to new storage every 3-5 years.

This implies a massive copy operation, which takes several months to complete at present. (Ref: 6, 7, 8, 21).



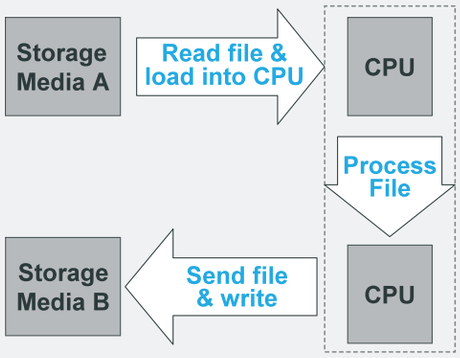
The Model of the Digital Repository at the National Library of Norway

MIGRATION TIME CALCULATOR

The National Library of Norway has 1 PB of genuine digital data (2008) and after digitalizing of all their records the data volume will be around 40 PB (ca. 570 million files). The support by storage vendors is limited to 3-4 years only requiring the migration of the entire data volume to new storage discs. It is easily envisioned that the previous migration is still not finished when a new migration has to start.

The LongRec project used a process modeling approach to derive formulas for the expected migration time. Processes covered by our model are: migration, file processing, replication and verification.

Test experiments at the National Library of Norway and NTNU indicated that the selection criteria for storage discs should not be the best or average write speed but rather the worst read / write performance. Also, verification does not increase migration time in multiprocessor systems. Interestingly, multiple PCs can obtain the same performance as advanced and expensive storage systems. In this respect, the methodology derived may be advantageous in migration planning and in selection of new storage hardware (Ref: 6, 7, 8).



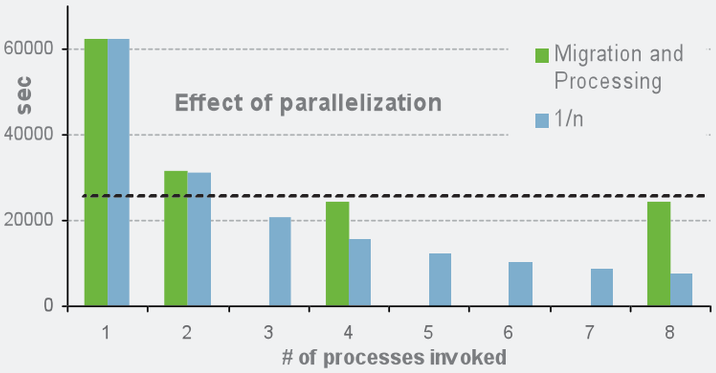
Based on a processing modeling approach a mathematical framework was derived for calculating the migration time.

THE POWER OF PARALLELIZATION

The National Library of Norway plans to convert about 12 mill files (ca 7500 TB) from tiff-format to jgp2000 format including a color correction requiring 5 s/file. This conversion alone would need nearly 700 days.

In the LongRec project conversion experiments were performed to examine the effect of utilizing parallel or multi-core processors. The results indicate that when using at least 9 parallel processors the conversion project could be finished in just about 80 days.

A high degree of parallelization can keep conversion and verification time at an acceptable level, but it has little effect on file migration, i.e. the write time of the converted file to its storage disc will be the time limiting factor (Ref: 6, 7, 8).

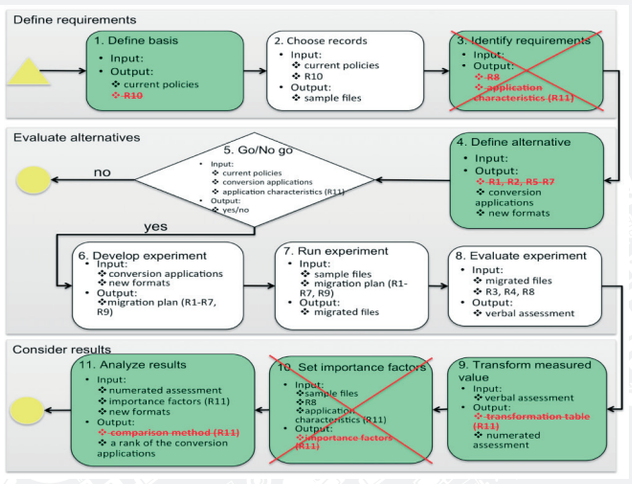


File conversion time drops with an increasing number of CPUs until the limits of the CPU is reached (blue dotted line) as background processes steal processing power.

QUALITY REQUIREMENTS OF MIGRATION METADATA

Migration metadata are essential but few requirements on these metadata are specified. A set of quality requirements for such migration metadata were derived from the commonly used metadata in OAIS, NLNZ and PREMIS. The quality requirements address 11 aspects covering hardware, application, specification, object, and policy.

To validate the completeness and the usefulness of these quality requirements they were applied to a workflow planning tool commonly used by many European libraries. The results show that the quality requirements indeed can play a very important role as they simplify and reduce the workload in a number of tasks. (Ref: 11)



Quality requirements on metadata can simplify the workflow for format selection in a planned migration.

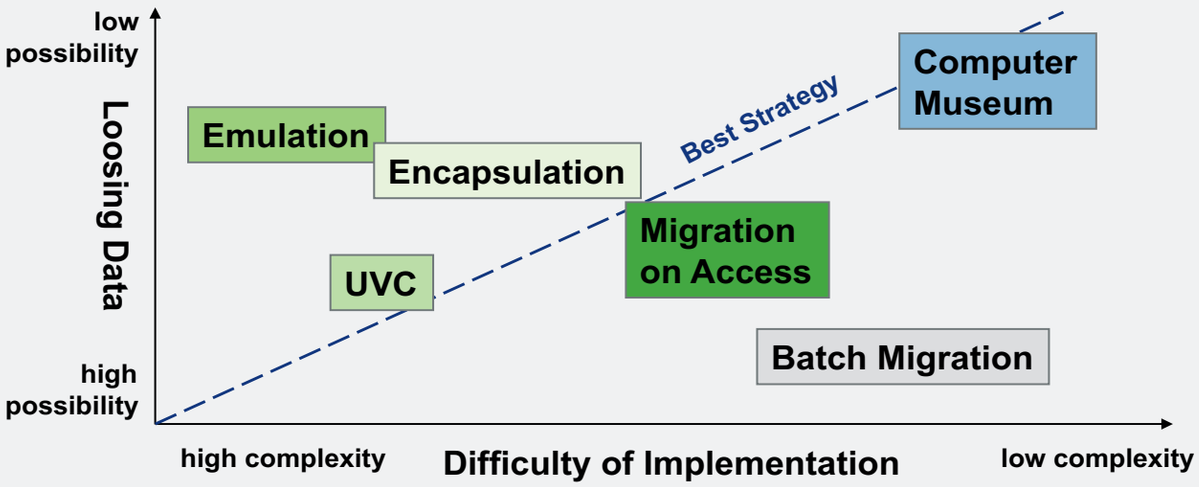
SURVEY OF PRESERVATION STRATEGIES

Digital information is made up of a set of unique objects which depend on the computer's help to render. The dependence on computers makes it difficult to preserve them over a long period since there are many threats, e.g., hardware/software failure, hardware/software obsolescence, loss of format specification, malicious modification, loss of context and organizational failures.

Many institutions have noticed this preservation challenge. Seven preservation strategies have been proposed: computer museum, emulation, Universal Virtual Computer (UVC), encapsulation, migration on ingest, migration on access and batch migration. Among those preservation strategies, computer museum and migration on ingest are treated as short-term solutions, while the remaining ones are long-term solutions.

For the long-term solutions, we design a matrix to compare them. The matrix contains six criteria:

- **"Change in bits"** means how many times the bits of the digital object have been changed over its lifetime.
- **"Hardware specification"** means whether hardware specifications should be preserved.
- **"Hardware components"** means whether hardware components should be preserved.
- **"Format specification"** means what format specifications we should preserve.
- **"Format converter"** means what format converter we need.
- **"Format interpreter"** means what kind of interpreter programs we need to translate formats *(Ref: 8, 21, 56).*

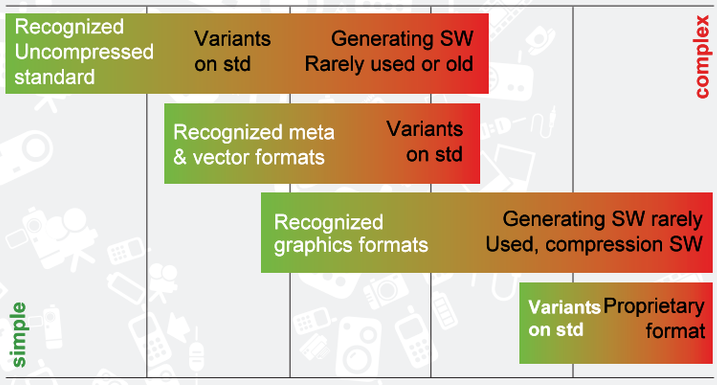


Losing data is determined by change in bits, format converter and format interpreter, while difficulty of implementation is determined by hardware specification, format specification, format converter, and format interpreter.

COMPLEX OBJECT = COMPLEX PRESERVATION

As there is no single best way of long-term preservation that fits all digital objects, it will have to be decided on a case to case basis what strategy should be chosen. This strategy will also have to be revised regularly.

In collaboration with LongRec an evaluation and decision support framework was developed to assist the user in selecting a favorable preservation strategy. This framework is based on a company's information quality requirements, general theory around information management and IT practices, and the various complexity of the digital objects. It visualizes storage options, helps to define processes for ordering and submission of databases for archiving. The framework was empirically validated. *(Ref: 8, 9, 57)*

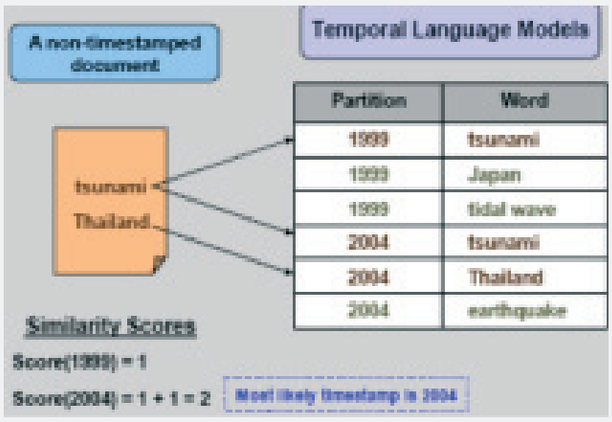


Object complexity is a major factor when selecting an appropriate preservation strategy.


ENHANCED PERFORMANCE DISKETTES
2/4
2HD
1.44 MB
MF 2HD

TEMPORAL LANGUAGE MODELS FOR DOCUMENT DATING:

In order to increase precision in searching for web pages or web documents, taking the temporal dimension into account is gaining increased interest. A particular problem for web documents found on the Internet is that, in general, no trustworthy timestamp is available. This is due to its decentralized nature and the lack of standards for time and date. For a given document with unknown timestamp, it is possible to find the time partition that mostly overlaps in term usage with the document. For example, if the document contains the word "tsunami" and corpus statistics shows this word was very frequently used in 2004/2005, it can be assumed that this time period is a good candidate for the document timestamp. The model assigns a probability to a document according to word statistics over time. *(Ref: 27)*

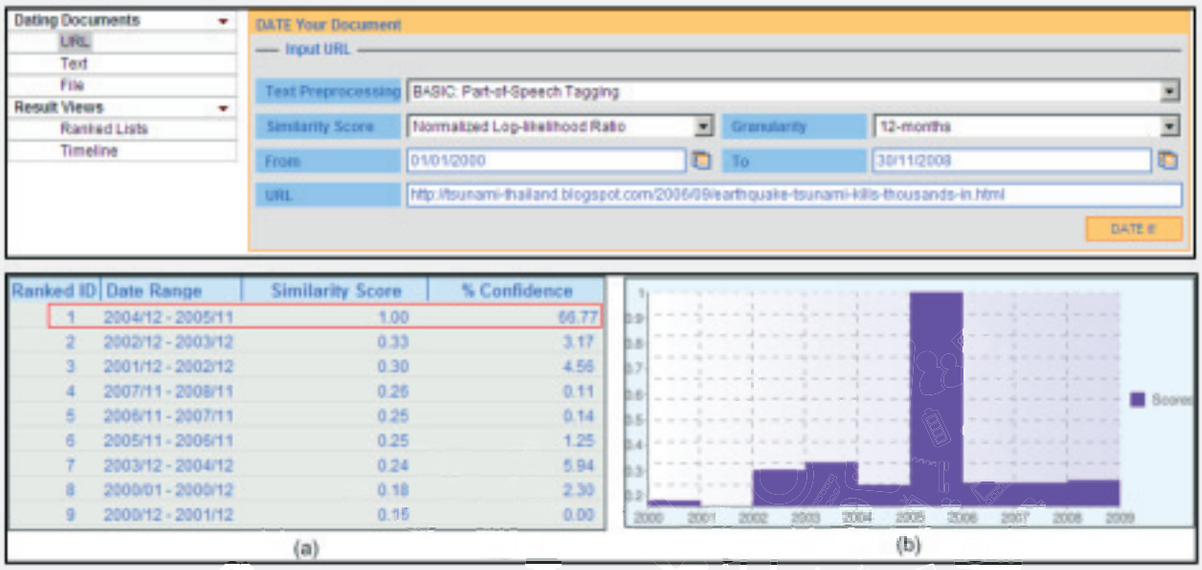


Temporal language models are based on the statistic usage of words over time.

DOCUMENT DATING TOOL:

A tool for determining timestamp of a non-time stamped document (i.e., a file, an URL or text as input) was developed using temporal language models. An example input can be the URL "http://tsunami-thailand.blogspot.com". A user can select parameters: preprocessing (POS, COLL, WSD, or CON), similarity score (NLLR, GZ or TE), and time granularity (1- month, 3-months, 6-months, or 12-months).

Given an input to be dated, the system computes similarity scores between a given document/text and temporal language models. The document is then associated with tentative time partitions or its likely originated timestamps. The results can be displayed in two ways. First, a rank list of partitions is shown in an descending order according to their scores. Second, each tentative time partition is drawn in a timeline with its score as a height. (Ref: 28)



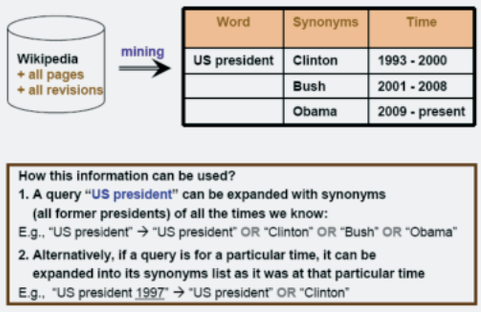
Document dating online demo at **http://comidor06.idi.ntnu.no:8080/timedelivery**

HANDLING SEMANTIC GAPS IN SEARCHING WEB ARCHIVES: THE USE OF TIME-DEPENDENT SYNONYMS

Web archiving is gaining interest as a way to preserve humanity knowledge for the future. Examples of web archive projects are: a large-scale web archive on the Internet Archive, newspaper archives of Google or the Times Online, and national web archives of national libraries.

One problem of searching such a document collection is the effect of rapidly changing synonyms of named entities over time, e.g., changes of roles or alterations of names. In order to retrieve relevant documents with respect to a particular time period, we need to handle these changing synonyms, called time-dependent synonyms.

We propose an approach to extract synonyms of named entities over time from a whole history of Wikipedia. In addition, we discuss how to make use of the time-dependent synonyms in order to increase recall, i.e., expanding a query with a set of synonym corresponding to a particular time period. to increase recall. *(Ref: 29).*



A dictionary of time-dependent synonyms can be used for query expansion.

QUEST: QUERY EXPANSION USING SYNONYMS OVER TIME

QUEST (Query Expansion using Synonyms over Time) is a system that exploits changing synonyms over time in searching news archives. This system consists of two parts: 1) the offline module for extracting time-based synonyms as depicted in Fig. 1, and 2) the online module for searching news archive as illustrated in Fig. 2. With a web-based interface, the system can take as input a named entity query. It automatically determines time-based synonyms for a given named entity, and ranks the synonyms by their time-based scores. Then, a user can expand the named entity with the synonyms in order to improve the retrieval effectiveness.

The time-based synonyms extracted using our approach can be applied to any news archive collection. In this demo, we use the New York Times Annotated Corpus as an illustrative example of such a news archive. This collection contains over 1.8 million articles from January 1987 to June 2007. We use the enterprise search platform Solr from Apache Lucene. *(Ref: 30),*

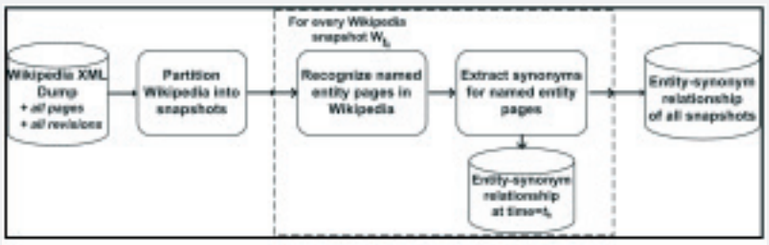


Fig. 1. Extracting time-based synonyms from the history of Wikipedia

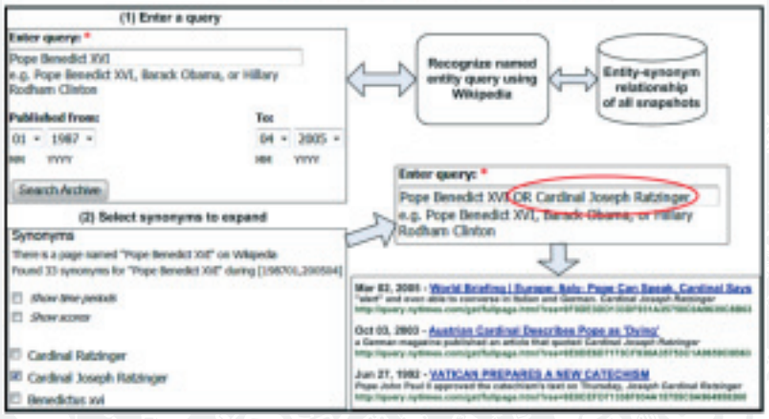


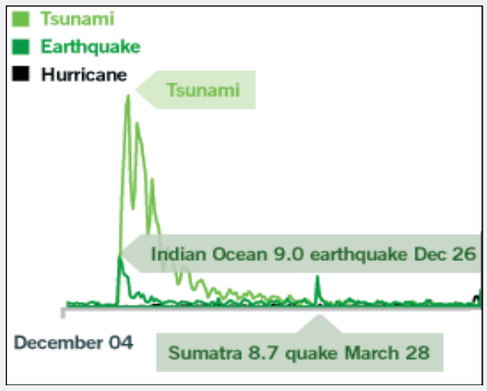
Fig. 2. QUEST online demo at http://research.idi.ntnu.no/wislab/quest

TEMPORAL RANKING:

In general, when searching news archives, hit-list documents are displayed in a chronological order where newer pages are more important/relevant than older ones. However, a chronological filtering is not always needed. Therefore, ranking the result documents by taking into account temporal information is necessary.

We propose to analyze a document collection to obtain a topical trend (the trend of a topic) that can be represented as the weight of a topic over time. For a given query, documents will be retrieved based on their similarity scores, e.g., TF-IDF to a query topic.

However, the ranking of documents is the combination of their similarity scores and document weights with respect to a topical trend. For more details about problems and possible solutions on searching document archives see *(Ref. 31)*.



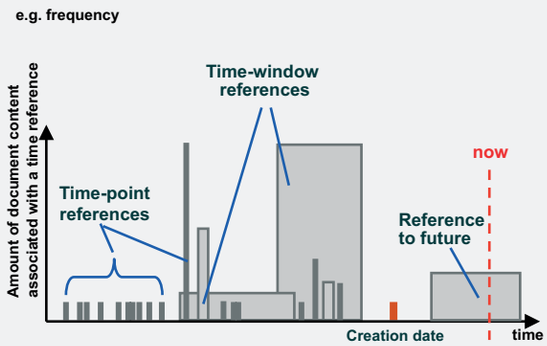
Trends of tsunami, earthquake and hurricane.
From Google zeitgeist (www.google.com/zeitgeist)

WHAT TIME IS IT?

It is said that 'time stands still' when nothing happens, and 'time flies' when a lot of changes occur. The effect of time on records is usually neglected as search algorithms do not account for semantic changes in the language. Neither has been much attention on visualizing the temporal aspects of a record.

During the LongRec project attention was directed towards suitable time references either in the form of time-points, e.g. date of birth, incident, or time-windows, e.g. WW2, fashion.

In order to counteract the effect of time the series of measures are available: storage of old search indices, logs of semantic changes, analysis of semantic changes, translation of new terms into old terms, and extraction of time references from meta data or the document content. A markup language for temporal and event expressions is under development (ISO/CD 24617-1). *(Ref: 3, 34)*



A fictive example of visualization of time references in a document. Thin lines refer to time points and rectangles to time windows.

IS IT POSSIBLE TO TUNE TO COMPLETE AND RELEVANT RANKING?

The case study addresses the challenges faced by the National Library of Norway regarding search and retrieval performed by the library's relatively new (2007) general search solution based on FAST technology. The quality of search in huge data volumes contained primarily in the National Library's trusted digital repository is the scope of the case. Search is based on indexing the structured information like bibliographic catalogues and unstructured information like OCR output in scanning processes. In this study it is proposed that the search quality may be measured in such terms as completeness and relevance of the search results.

The main research question is whether the quality of hits returned by the search engine can be improved after conducting the analysis and subsequent tuning of the built-in ranking algorithms and client profiling. *(Ref: 2)*.

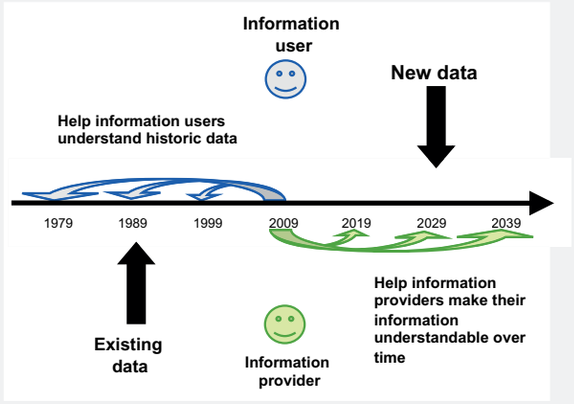
The GUI of the Search Engine at the Web-site of the National Library of Norway, www.nb.no


01:)101
01010

PRESERVATION OF SEMANTIC VALUE – THE STUDY OF THE STATE OF THE ART

The purpose of the state of the art report 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 which are: (i) the symbols/words used, (ii) the meaning of symbols/words and (iii) the domain knowledge. An extensive literature study has been performed in areas such as records preservation, library science, literature science, semantic technology and others. Literature on "semantic preservation" is somewhat limited, and we therefore believe that this is a somewhat immature area.

The state of the art report list several causes for changes in semantic value/ meaning e.g. change in symbol/term, reference, referent, record, worldview and counter-measures. These topics are described and discussed in the report, together with a review of current best practice. *(Ref: 4, 53, 54)*



UTILIZING AGING MASTER DATA AT THE BRØNNØYSUND REGISTER CENTRE (PILOT AND CASE STUDY)

The main challenge in this case study is the lack of tool support for interpreting existing master data (Enterprise data at the Brønnøysund Register Center) in its historic context. This knowledge is silent knowledge residing in the heads of senior employees. There is no information governance policy focusing on leveraging the implicit semantics of Business Enterprise data for the future.

The screenshot from the pilot illustrates how master data are linked visually along the timeline as an alignment anchor between several data sources needed for optimal utilization of the historical versions of master data. The usefulness of the pilot was tested on a small panel of users.
The pilot used a simple ontology, some semantic technology software and principles from the Semantic Web - Linked Open Data initiative. *(Ref: 5, 12, 52)*

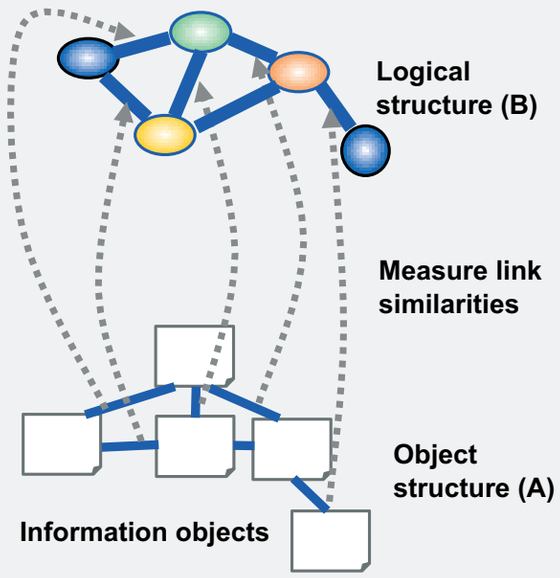
VALIDATION OF LOGIC IN THE INFORMATION STRUCTURE

Traditionally we have been using logical structures in web page hierarchies of web portals and filing key of documents in archives. Both of these structures are usually man made, and a large number of web-pages or records are linked to these structures. In large structures of this kind it becomes hard to validate whether or not the structure actually is at the quality level needed.

To help evaluate the quality we made an test case based on DNV web pages. In general term what we did was the following:

- There exists a structure (A) between objects
- Automated discovery and extraction of logical structure (B) from the website
- Automated validation of similarities between structure A and B

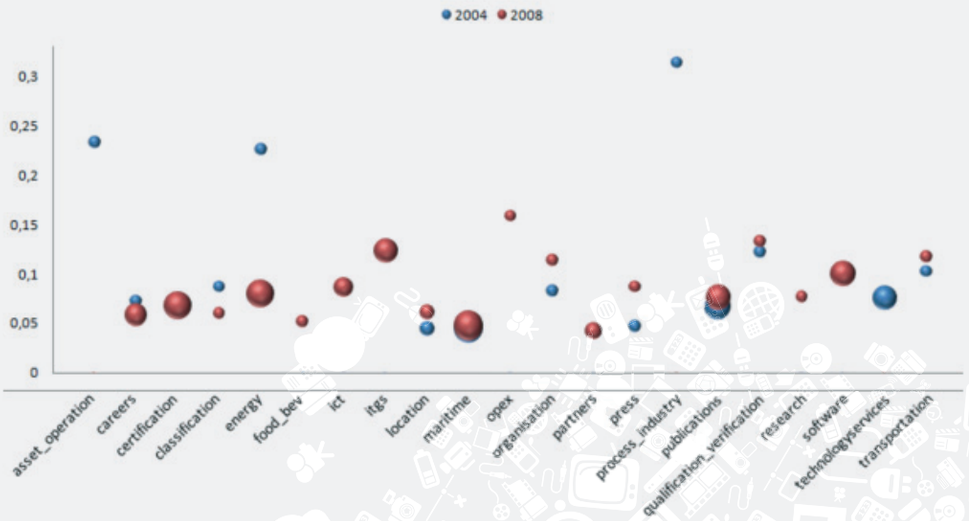
One of the rules we tested was sub-super relationships. E.g. that a child web-page should be stronger semantically linked to its mother web-page than to its sibling web-pages. In the DNV web page hierarchy the calculated sub-super relationship had god similarities with the man made structure of the web portal. *(Ref: 18, 19)*



DETECTING CHANGES IN THE USAGE OF TERMINOLOGY (MEANING OF TERMS)

Semantic drift refers to how the meaning of terms, the concept, gradually change as the domain evolves. When semantic drift is detected, it means that the concept is gradually understood and used in a different way or its relationships with other concepts have been undergoing some changes. This is an indicator for changes in knowledge or ontology that captures small domain changes that are hard to detect with traditional approaches like ontology engineering or ontology learning.

We make use of concept signatures as calculated vectors. The vectors are constructed on the basis of how concepts are used and described in text. Comparing how signatures change over time, we see how concepts semantic content evolves and how their relationships to other concepts gradually reflect these changes. An experiment with the changes at the DNV web-sites from 2004 and 2008 demonstrates the value of the approach in ontology evolution. *(Ref: 20)*



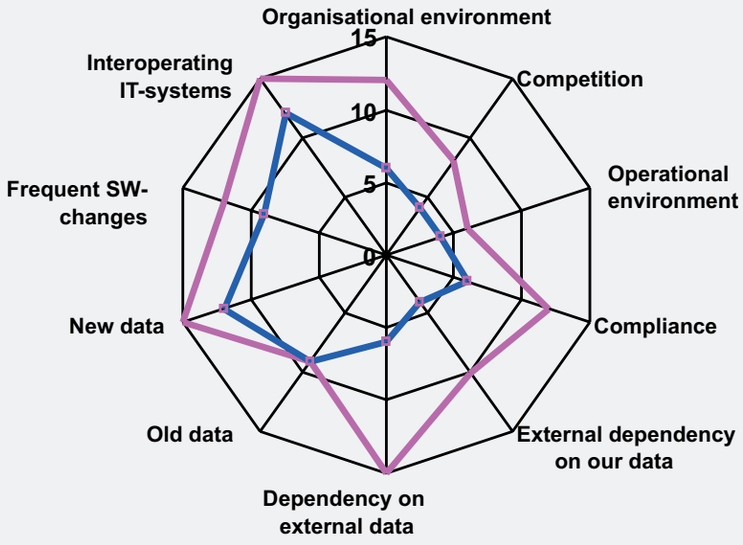
Frequency in usage (vertical axis) of terms (horizontal axis) related to the DNV Consulting web site between 2004 (blue) and 2008 (red bullets). Bullets close to each other indicate stable concepts with respect to the reference term Consulting.

INFORMATION GOVERNANCE REGIME (BASED ON THE LONGREC UNDERSTAND CASE STUDY)

The Longrec Understand case partner wanted to improve their long term information governance by utilizing semantic methodologies and tools. To achieve this methodologies were applied that automatically calculate and measure hierarchies of concepts and thereby detect an occurring semantic drift. Legal aspects of mashups were also discussed as they will increasingly have to be taken into consideration.

To validate the chosen approach and to gain necessary experience a pilot or prototype (Enterprise History Interpreter) was developed. In order to improve the longterm information governance the following topics were specifically addressed *(Ref: 4).*

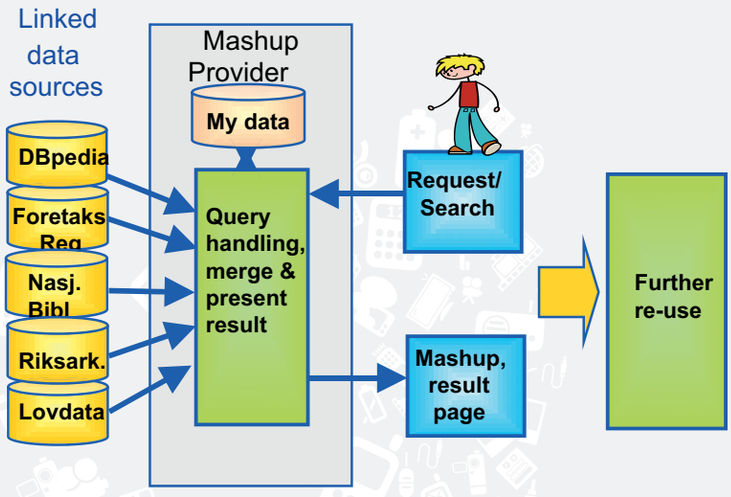
1. Management survey, to assure management focus on information governance.
2. Issues related to compliance in information governance.
3. A list of existing and alternative information governance regimes.
4. A brief process of how to establish an information governance regime and the use of the management survey as part of this process.



LEGAL ASPECTS OF MASHUPS

In web development, a mashup is a web page or application that combines data or functionality from two or more sources to create a new service. Based on the experience from the Longrec Understand Pilot, there was a need to asses legal aspects related both to the mashup in the pilot and a potential third-party-mashup using data from the registers at the Brønnøysund Register Centre. The following topics are discussed:

- Protecting own rights and needs as a mashup and information provider
- Protecting rights of 3rd parties (sub contractors / information providers), in "my" mashup.
- Limiting own legal responsibility



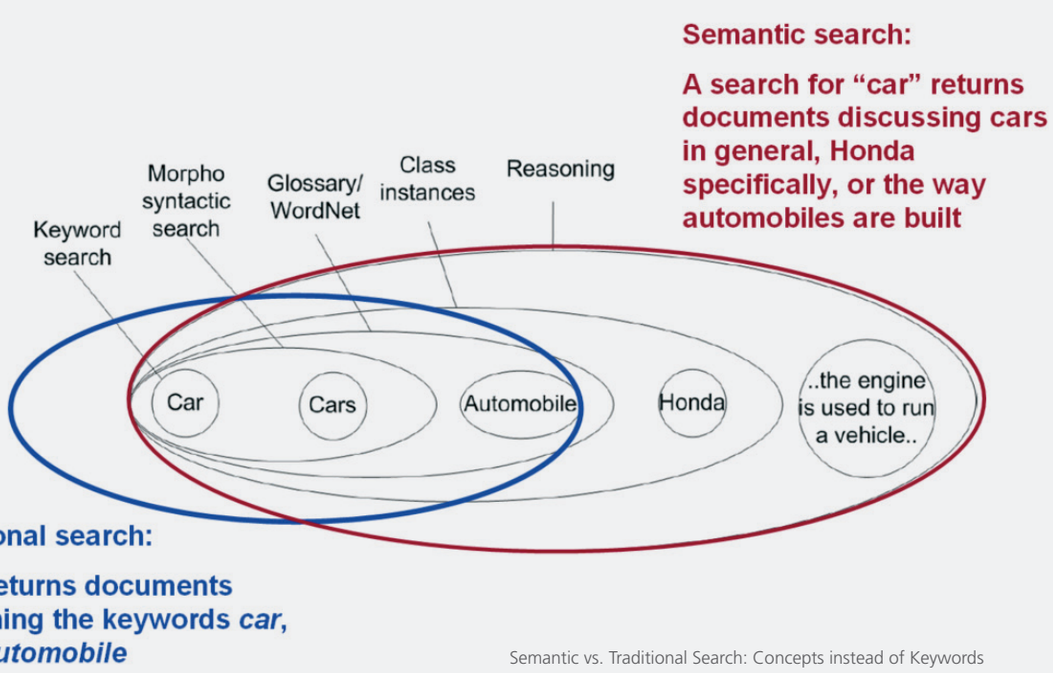
The topic is generally valid for data on the web, and especially for public sector information initiatives and the Semantic Web - Linked open data concepts *(Ref: 4).*

SEMANTICS AND SEARCH

The sheer amount of information is one of the most challenging aspects of Internet and enterprise search applications, but there are also other aspects that hamper the effectiveness of current search technology. Standard search applications' reliance on simple keywords is satisfactorily for users that know exactly which words should appear in the documents they request. However, most users will not know in detail the wording of all documents, and the flexibility of natural languages makes it difficult to guess how words and phrases are used to describe phenomena. If terminologies also change over time, a suitable keyword today may fail to identify relevant documents from the past.

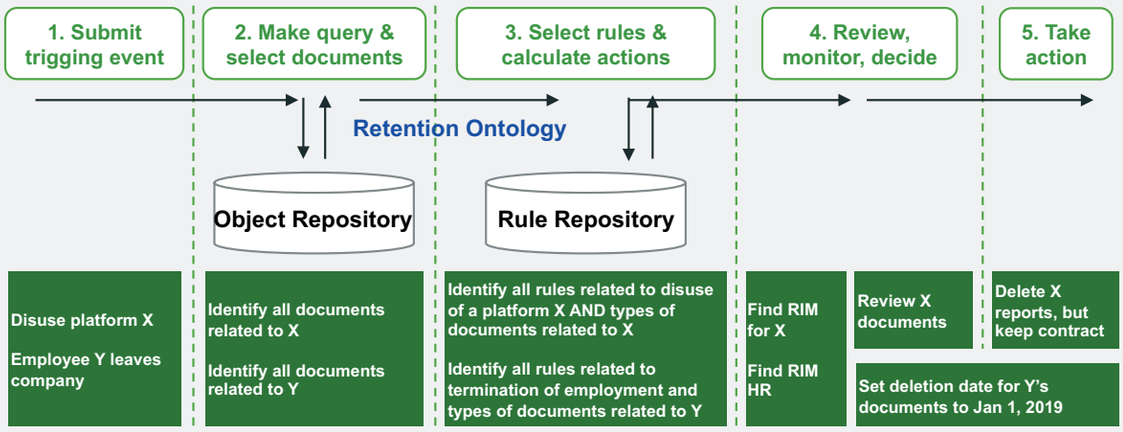
Enterprises observe the increasing importance of good search facilities for managing their internal knowledge and resources. To handle the complexity of their businesses and take full advantage of their own competence, they need appropriate tools for documenting and retrieving business-critical information. This is a particular concern in evolving domains, in which organizations are constantly changing and need to relate to new procedures, new technology and new staff.

Semantic search applications address this language problem of current search technology. They offer mechanisms for dealing with document content rather than keywords, and they try to capture the variety and instability of terminologies used in documents and queries *(Ref: 22)*



Semantic vs. Traditional Search: Concepts instead of Keywords

AUTOMATIC RETENTION MANAGEMENT ENGINE



Information retention – especially of electronic data – has become a hot topic in the legal domain. In the wake of several court decisions leading to high-dollar jury verdicts, companies need to have good retention and preservation practices in place. 99% of all documents are created and stored electronically and somewhere around 60 billion e-mails are created and sent daily according to IDC. Electronic information is not only found on laptops; it is captured in PDAs, mobile phones, i-pods, and the list of storage media continues to grow. In the long-term perspective the picture gets quite complex as the organisations change over time and with this also the retention and disposition rules alter.

According to the laws and regulations it is as important to delete information when required as to preserve it.

The solution to this is to integrate a retention management engine with the clients' infrastructure. That will automatically update the settings related to retention. We have used the semantic technology solutions in combination with the architecture, business and compliance competence to arrive at the solution in this area. Companies can reduce their costs related to data management by using this engine integrated with their architecture *(Ref: 26, 40).*

COMPLIANCE STATE-OF-THE-ART

This document addresses a number of themes in compliance and records management domains:

Legal informatics studies the application of information technology to the practice of the law. Descriptions of concepts and the links between them alone are not enough to allow for automated processing of legal information. A method to express legal rules or legal logic in machine readable form is required. This is being researched in a number of projects.

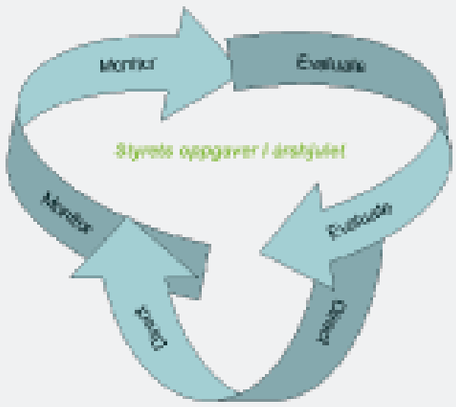
Electronic evidence has challenged existing legal practices with respect to collection, production and evaluation of proof. In the discussion of electronic evidence, two distinct legal domains can be discerned, one is cyber-crime and the other is electronic transactions in a broad sense. Though electronic records management, as a business process, falls within the scope of the latter, insights gained in the former can be instructive. A trustworthy record management system is, therefore, one that can be relied upon to provide irrefutable evidence of all of the events that have been logged.

Digital technology has provoked an renewed interest in the probative value of copies of original documents. In many countries, the probative value of copies is dealt with in a piecemeal fashion

Additional themes include but are not limited to: Technological tools for capturing and handling evidence , Legal Metadata *(Ref: 35).*

HOW TO GET THE NORWEGIAN CEOs SLEEP WELL AT NIGHT – NORSOX

Norway needs to adapt the EU-directive (4,7 and 8). Unofficially this has been called EuroSox (the European SOX). Staying compliant is the main challenge here. Standard Norge invited several companies to join the NorSOX research project. DNV has been part of this project both from the business side and the research side (LongRec). The main recommendation is that the the board evaluates the it-use/information asset within the company related tot the 4,7 and 8 EU-directive in a designated meeting at least once a year. The results of this review should be reflected in the corporate business plan. A way of doing this is by using the wheel within the ISO 38500. There are 3 steps: Evaluate, Direct and Monitor For each of this steps there is a checklist to be considered by the board.
(Ref: 37, 38)

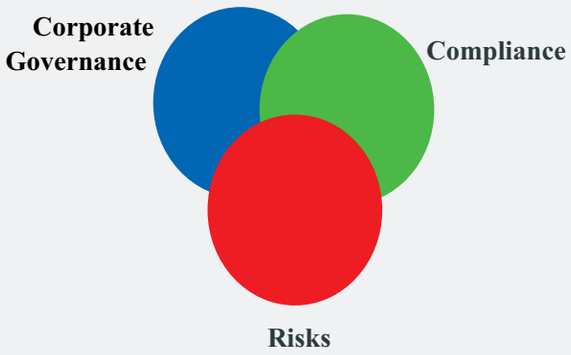


The tasks of the board related to "NorSOX".
("Norway as a EEA member will have to introduce three EU-directives into Norwegian law, 4, 7. and 8. company directive.")

HOW TO ENSURE YOUR CEO'S GOOD-NIGHT SLEEP – COMPLIANCE TOOLBOX

The companies participating in LongRec as well as other companies inquire how they should handle compliances issues related to their information assets in order to prevent scandals like that with Enron. LongRec has tried to help the partners with this problem by studying the partners' needs as well as the suggestions from several standardization bodies. The solutions to this is to first let the board decide what risks the company is willing to take regarding the information assets and then to identify how information management should be handled within this scope. After that the follow-up by using the maturity model is carried out. The toolbox contains 3 steps:
1. Corporate Governance
2. 10 Steps Information Governance
3. Information Maturity Assessment

(Ref: 36, 42)

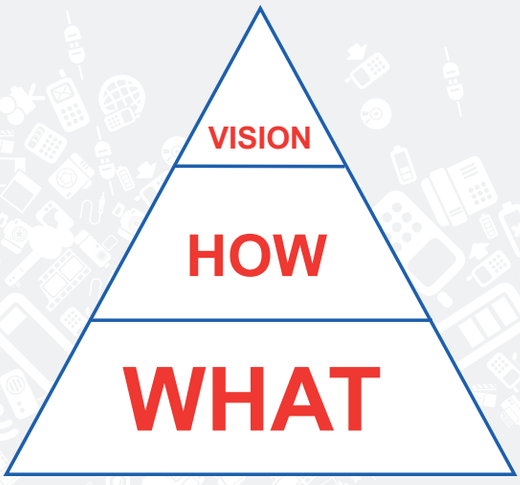


You identify what kind of laws/regulations you will prioritize related to information governance and by doing this you also choose the relevant risk level related to compliance issues.

COMPANY GUIDELINES RELATED TO RETENTION

Regulatory demands and the number of documents produced daily continue to grow. Companies needs to preserve electronic information only as long as necessary for business purposes (history, compliance and operation). Therefore, a solid document management process is a necessity for proper retention. This should be collected within a retention policy.

The information assets need to be broken down to the least information objects and for each of them one will need to identify: the Retention rules (if you have a conflict related to retention time which one should you follow), retention time/period, work process, disposition rules, information owners, types of applications, types of servers, types of media, how the backup is handled and so on. All of this should be specified in an retention policy. ISO 15489 is a key guiding instrument for professional record managers in this topic. *(Ref: 26, 36, 42, 55)*



The retention policy needs to be linked with the business needs.

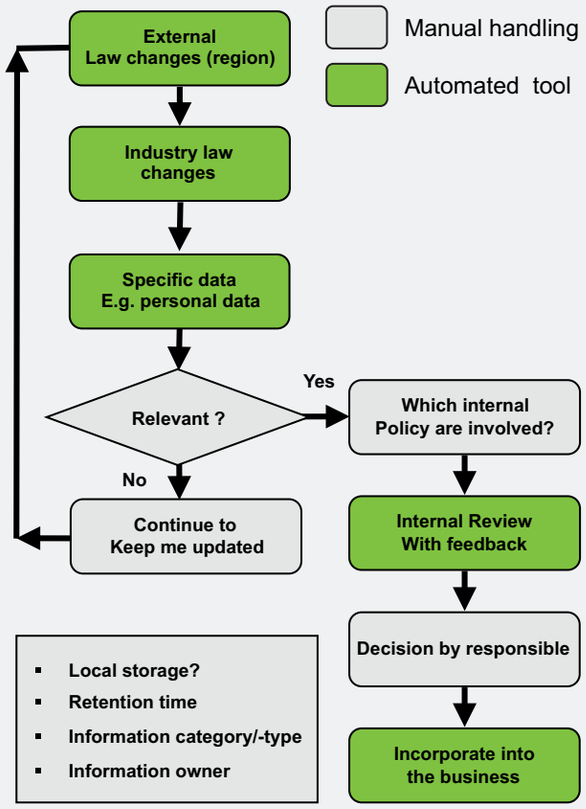
COMPLIANCE TOOL

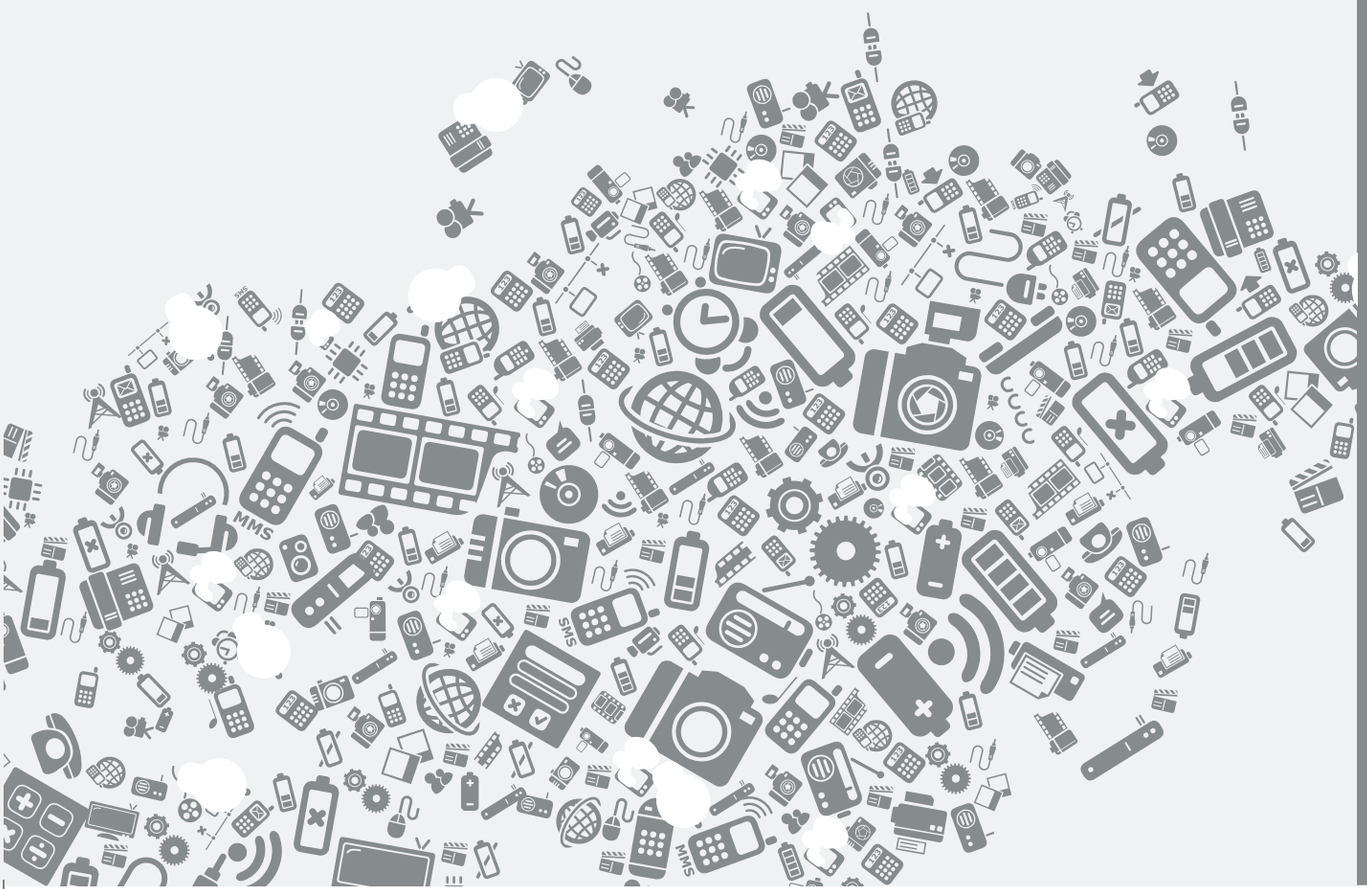
The reason for long term storage of data are often linked to compliance issues, either company internal or external. E.g. in Norway the laws; "arkivloven, offentlighetsloven, regnskapsloven, personalopplysningsloven, helselov-givningen, sikkerhetslovgivningen etc" all affect information management issues. In a global company like DNV or Statoil which are represented in several countries with their different laws/regulations, information management which in addition will change over time. The vision is to get a compliance tool that gives an alert whenever a law/regulation changes and offer guidance through the required changes within your company. Ideally, such a tool should be fully automated or should be combined with a manual workflow. The LongRec project has approached the Rettsinformatikk for a feasibility assessment of such a tool.

Response from Rettsinformatikk v/Jon Bing National (Ref: 13)
Lovdata has confirmed that they could together with a client develop such a tool and offer an alert service related to information management.

International:
Globally, there are two similar sources: WorldLII and GLN, but none of them can guarantee complete coverage of a country.

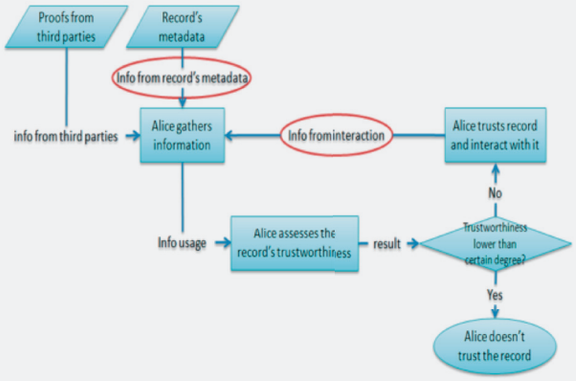
The challenges are in general:
- Language
- Such a tool requires that (all) countries must have a change alert system in place.
- International services like WorldLII and GLN will never be better than the available national services.
- the organization and maintenance of laws/regulations are different from country to country. More research would be required.





THE TRUST MODEL

The trust model in the real world:
When Alice meets Bob for the first time, she assesses Bob's trustworthiness by his behavior and/or information from other people who might know Bob from before. If Alice decides to interact (by chatting, cooperating, etc) with Bob, further trustworthiness assessment will mainly be based on the information collected from these interactions.
(Ref. 15, 16)

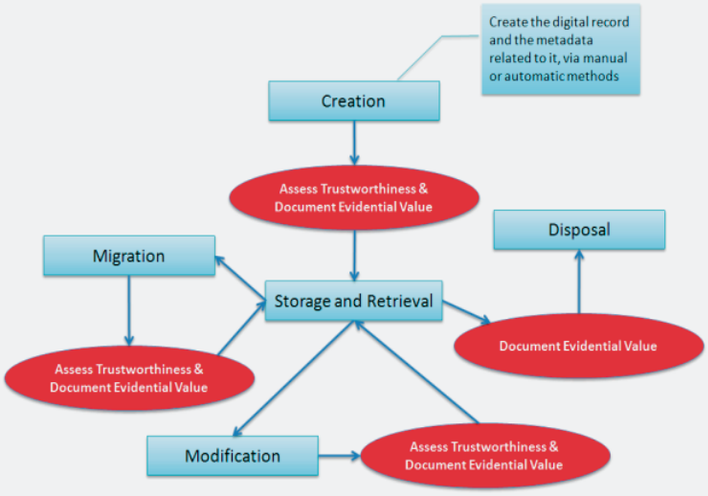


The trust model for digital records:
When Alice encounters a digital record for the first time, she assesses the record's trustworthiness by evaluating the metadata of the record such as record's name, creator, creation date, third parties' proof (digital signature), etc. If Alice decides to interact (by reading, editing, etc.) with the record, the subsequent trustworthiness assessment will be based on the information which was collected during these interactions and will be documented in the metadata contributing to the overall evidential value. *(Ref. 15, 16)*

REQUIREMENT FOR TRUSTWORTHINESS ASSESSMENT:

The digital records in long-term repositories are intended to be preserved for many decades. However, with the current technology, it might be hard to evaluate whether a record from the repository is trustworthy or not. By comparing trust in the real world with trust in the digital world, the LongRec project illustrated how and why evidential value is essential in the assessment of the trustworthiness of a record.

We proposed a record's life cycle model which give much attention to those phases during which the trustworthiness of a record is likely to be compromised, and little attention to those phase during which it will not be compromised. We identified, analyzed and specified the requirements for evidential value for the assessment of the trustworthiness of digital records at each phase.
We also validated these requirements by performing a web-based questionnaire survey. *(Ref. 15, 16)*

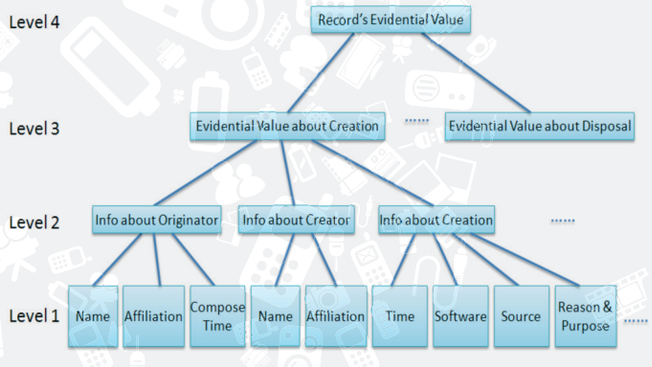


APPROACH TO ASSESS THE RECORD'S TRUSTWORTHINESS

Approach to Assess the Record's Trustworthiness
Currently, there is a lack of approach to calculate the trustworthiness of digital records. By identifying the requirements for evidential value and structuring them into a tree structure, the LongRec project will develop an approach to assess the trustworthiness of digital records over time.

This assessment approach calculates the trustworthiness of a record's sub-attributes from the bottom (level 1), and then combine the results from these sub-attributes to calculate the trustworthiness in a higher level. Finally, the trustworthiness of the digital record can be received from the top level of this tree structure.

The archival organization can use this approach to convince their customers about the trustworthiness of the digital record *(Ref. 17)*

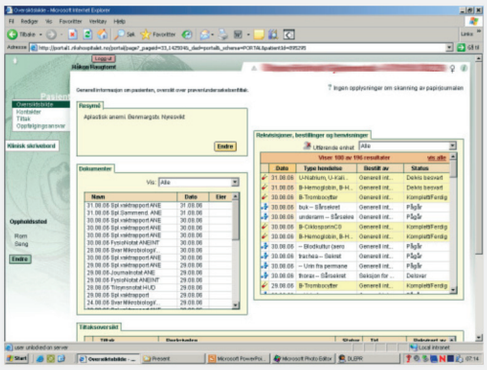


The tree structure of a record's evidential value

CORRECT AND COMPLETE ELECTRONIC PATIENT RECORDS IN A PORTAL SOLUTION

CSAM AS – Clinical Systems All Managed Ltd. – provides a portal solution that integrates different systems containing electronic medical records, achieving a component based electronic patient record. The main research focus was on how to preserve the correct presentation and completeness of clinical information across systems integrated by the portal in the long term.

Complicating factors of this case are the growing volume of digital patient information and its heterogeneity, which includes not only written material but also variety of formats and accompanying hardware, images, video files, etc. Transitions of the records between hardware and software as well as conversions to newer formats are inevitable in the course of life of a patient (Ref. 39, 59)



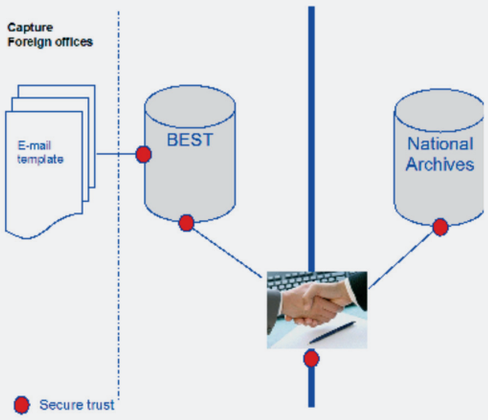
Can a user be certain that the patientdata depicted is complete and correct at all times

- How is information completeness secured in all the integrated systems and back to "time zero" for a patient?
- How can we reconstruct information that existed about a patient at a given time?
- How can we trust that the reconstructed information is correct?

TRUST IN PUBLIC CASE HANDLING AND ARCHIVAL TRANSFER OF RECORDS TO THE NATIONAL ARCHIVAL SERVICES

The LongRec case study has investigated the trust challenges associated with e-correspondence (e-mails) used in public case handling at the Ministry of Foreign Affairs (UD) and archival transfer to the National Archival Services of Norway (RA), (Ref: 23). This includes.

- How to capture e-mails in the internal records management and archive system preserving the adequate integrity and authentication information?
- How to confirm the sustained integrity of the archival information content after possible migrations and conversions?
- What is important to pay attention to when preparing the e-mail records for the transfer to the Archival Services?
- How to confirm that the archival version received by the Archival Services is identical to the version produced by the Ministry of Foreign Affairs?



The red dots indicate the points of trust in the workflow where the electronic records may potentially be compromised.

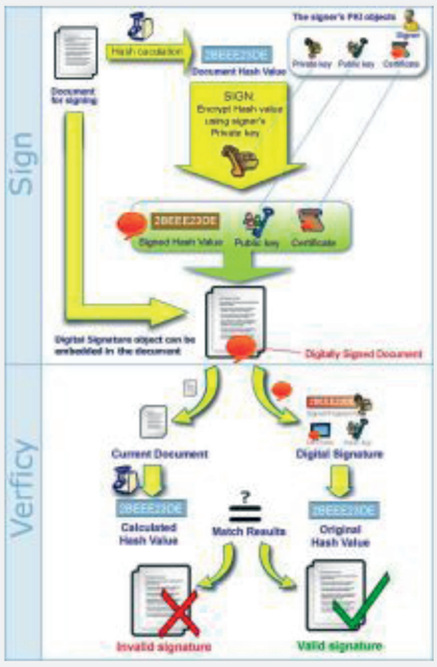
CAN ELECTRONIC AND DIGITAL SIGNATURES LAST FOREVER?

The main risk regarding e-signatures is the risk of forgery. Sufficiently strong cryptographic algorithms and sufficiently long keys are meant to alleviate this risk. However, over time the cryptographic strength will weaken.

Other long-term challenges related to the digital signatures are:

- Lifetime (expiry, revocation) of the keys and certificates used
- Lifetime of the signing method
- Lifetime of formats of content, signature, signed data object, certificate, and other supporting information like time-stamps
- Lifetime and continued service offer of (trusted and other) actors upon which the verification process relies

For the verification of a digital signature, an entire validation chain is necessary. It is not sufficient just to archive the document and the digital signature if one wants to validate a digital document with a digital signature in the future.
The value of the digital signature depends therefore on the procedure in which the digital signature is used, the association between private key and owner, and on the safe storage of the private key. Several preservation strategies are assessed in *(Ref: 1, 21, 60).*



Creation and Verification of Digital Signatures

HOW TO MAINTAIN ELECTRONIC SIGNATURES CREATED WITH BANKID?

E-signature maintenance is necessary in order to preserve its trustworthiness. Actors using BankID signing for legally binding agreements will be affected.

The first step for an actor using BankID is to identify whether technical maintenance of e-signatures is necessary within the outcome of 2011. Especially this concerns signatures created prior to the outcome of the year 2010. Risk assessment is the next step to determine the correct maintenance strategy for each set of records. This is then succeeded by the practical maintenance. From 2011 BankID Cooperation will offer a signature maintenance service
(Ref: 1, 60).

Two types of BankIDs prevail in the market: - stored in a bank or in a mobile phone. BBS (Bankenes Betalingssentral) is responsible for the storage of all Norwegian bank IDs.

TRUST STRATEGY BACKGROUND

Trust strategies can be classified along the following lines:

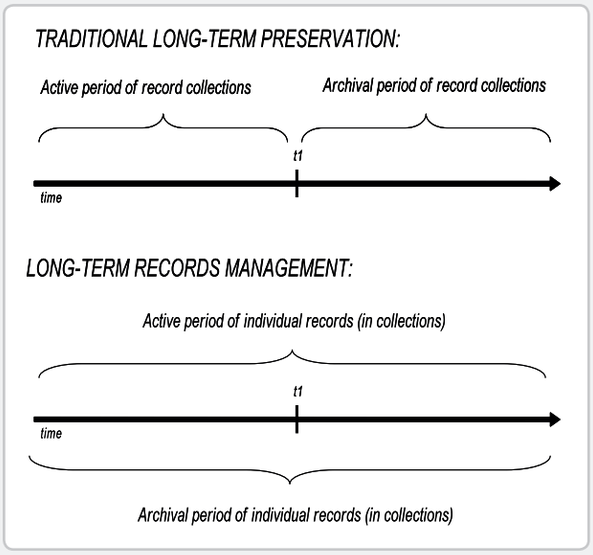
- One type of strategy is the **centralized strategy** and is based on believing in authorities. If an agent has a certificate of a certain authority, it can be trusted.
- In an **optimistic trust strategy** no additional justifications (evidence) is required. Optimistic trust strategies are sometimes the most cost-effective, under the precondition of a minimal risk.
- In the type of strategy called the **investigating strategy**, or the policy driven strategy, the uncertainty is reduced by investigating or evaluating details of other agents
- In "small-world trust", also called **pessimistic trust strategy**, one only trusts somebody or something for whom the previous history or relation of trust exists, i.e. the trust is based on (personal) experience. *(Ref: 10)*

LONG-TERM PRESERVATION VS. LONG-TERM RECORDS MANAGEMENT

In the traditional view of (digital) preservation/archiving, collections of (digital) records goes through an active period until they are no longer in business use. Then after a filtering period, the whole collection is ingested into the archive, often following a complex procedure due to the large number of records to be ingested.

If the active period of (digital) records become long-term, e.g. lasting several decades, challenges concerning archiving and preservation have to be handled during the active period of the record. Otherwise key information, evidence, and readability will be lost before the records become passive.

Either a dual world including two synchronized data representations of records, one archival representation and one records management representation, has to be created. Or, the active representation has to fit (minimum) archival criteria. Anyhow, archival ingest has to be considered at individual record level and not on collection level. *(Ref: 10)*



TRUST STRATEGIES FOR LONG-TERM PRESERVATION AND LONG-TERM RECORDS MANAGEMENT

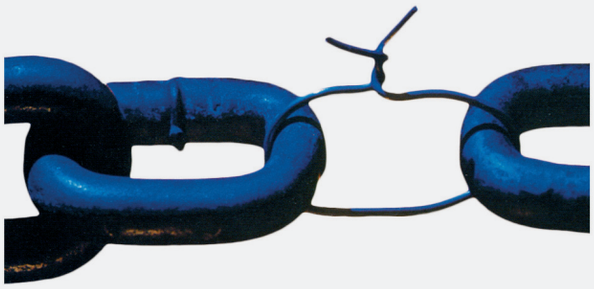
As part of the LongRec project various trust strategies have been studied in order to present an overview of the various trust strategy directions that can be followed, (Ref: 10),

including :

- Trust strategies based on diplomatic and forensic analysis
- Trust strategies for capturing provenance (changes to the material)
- Trust strategies based on encapsulation and cryptographic measures
- Trust strategies for validating digital signatures
- Redundancy used as a trust strategy
- Trust strategies for SQL databases.

TRUST STRATEGY: ARE THE RECORDS RELIABLE? – TRUSTWORTHINESS RIGHT FROM RECORD CREATION

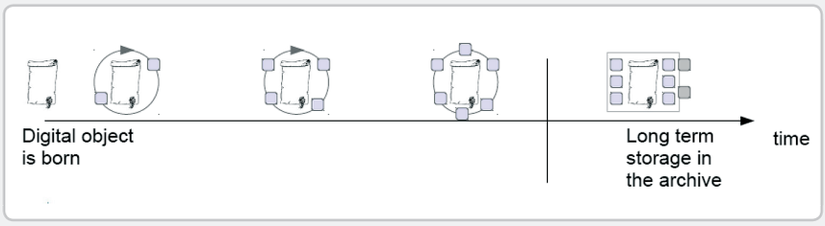
Information is captured to become a record as part of an organization's conduct of business (at record creation). The recorded information can be either an email, a digital document, a scanned document, digital video/audio etc. As part of creating records, the recorded information is annotated by various metadata.



To keep records trustworthy, various types of evidence should be collected and valued with respect to evidential value as part of the record creation process. To use an analogy form the non-digital world, a hand written, unsigned piece of paper is less reliable than an official signed document. The same applies to digital information that is about to become a digital record. *(Ref: 10)*

TRUST STRATEGY: DOCUMENTING EVERY ACTION ON THE RECORDS

Ideally, every action, by authorized actors, should be documented, together with the pre- and post state of the action. Unauthorized actors should not be put in the position of being able to modify anything.



Various audit trail mechanisms can be used for collecting the information, but the information is not useful and the evidential value decreases unless they are stored properly in secure storage from the time of collection.

Some sort of encryption mechanisms should also be used to secure the accumulated evidence. *(Ref: 10)*

TRUST STRATEGY: AUTHENTICITY THROUGH DESCRIPTION – BASED ON DIPLOMATIC ANALYSIS

The InterPARES 1 project has been using diplomatic analysis as one major tool for establishment of trustworthiness. They have produced two sets of requirements; The benchmark requirements forming a basis for presuming or verifying the authenticity of the creator's digital records; and The baseline requirements support the production of authentic copies of digital records after they have been transferred to the preserver's custody; Both sets of requirements define and give a basis for assessing the records' identity and integrity, which must be preserved for the copies to be authentic.

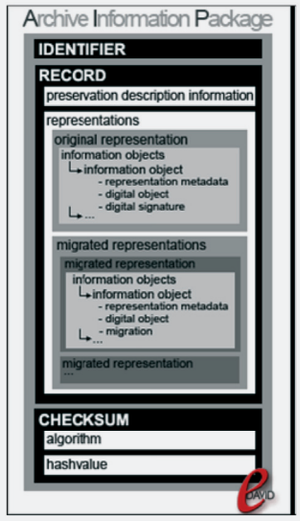
In addition, InterPARES 2 has developed a set of "Creators Guidelines", to be used by record creators as a tool for proactively being able to support reliability and authenticity right from the time of record creation. *(Ref: 10)*

TRUST STRATEGY: ENCAPSULATION

An electronic records management system represents the metadata in database tables, combined with file storage for digital documents, pictures, maps etc. In addition, SQL is full of constructs varying from one relational database implementation to another.

Therefore, a strategy of encapsulating data and metadata for long-term storage should be defined and followed during records management, preferably creating long-term storage representations on the fly during records management.

XML-encapsulation of both digital content and associated metadata has become state-of-art with respect to trustworthiness, in combination with cryptographic measures . To secure the integrity of entities, cryptographic hashes used as check sums in XML encapsulations should be used or even digital signatures if the digital records management system/the digital archival system support signing of digital material
(Ref: 10).



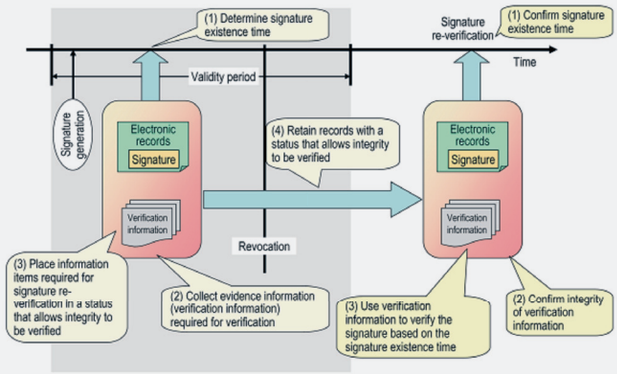
Encapsulation of the AIPs in the eDAVID approach.

TRUST STRATEGY: DEMONSTRATING FORMER VALIDITY OF DIGITAL SIGNATURES

There are two main approaches to demonstrate former validity of a digital signature:

- Documentation on e-signatures validity, or
- Ability to revalidate e-signature.

To preserve the long-term authenticity of electronic records the EVERSIGN approach proposes a solution they call Signature Validity Extension. Their solution makes use of the long-term signature format in the standard RFC3126. This is one of several approaches available. *(Ref: 10)*

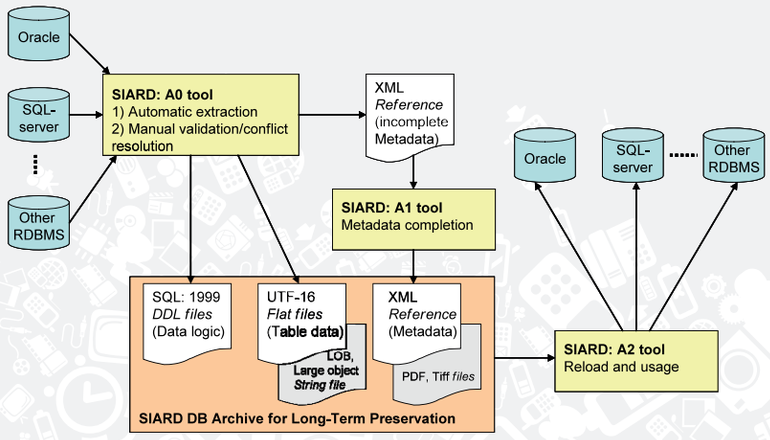


The EVERSIGN approach

TRUST STRATEGY: AUTHENTICITY THROUGH DESCRIPTION – BASED ON DIPLOMATIC ANALYSIS

The InterPARES 1 project has been using diplomatic analysis as one major tool for establishment of trustworthiness. They have produced two sets of requirements; The benchmark requirements forming a basis for presuming or verifying the authenticity of the creator's digital records; and The baseline requirements support the production of authentic copies of digital records after they have been transferred to the preserver's custody; Both sets of requirements define and give a basis for assessing the records' identity and integrity, which must be preserved for the copies to be authentic.

In addition, InterPARES 2 has developed a set of "Creators Guidelines", to be used by record creators as a tool for proactively being able to support reliability and authenticity right from the time of record creation. *(Ref: 10)*



The SIARD approach


КОПЕЕК
10

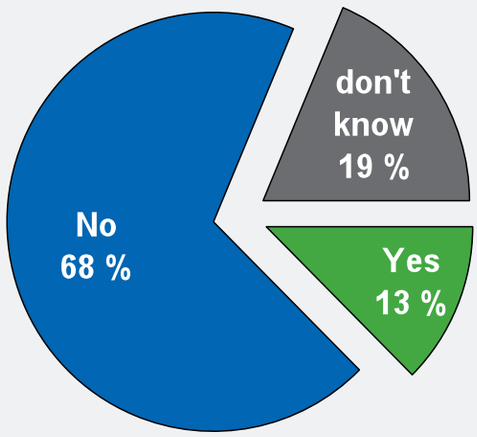
PRESERVATION COST MODEL

Reliable cost estimation is hampered by the constantly evolving technological environment leading to high uncertainties.

In collaboration with the LongRec, a M.Sc. thesis derived preservation-cost factors from the OAIS framework by decomposing its functional entities. These factors were validated by an empirical case study.

The study showed that almost 90% of the asked organizations either did not have or did not know if there exists a cost model. This lack of cost awareness is quite symptomatic for the long-term preservation. It became also apparent that resources were rather allocated on storage and technology than on human capital.

It remains to translate the identified cost factors to more commonly used monetary terms *(Ref: 25)*.

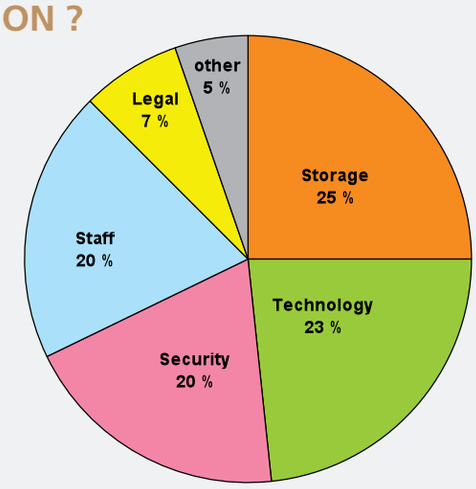


The empirical study showed that the vast majority of the respondents either didn't have or didn't know if their organization used a cost model in digital preservation.

WHERE ARE THE COSTS IN PRESERVATION ?

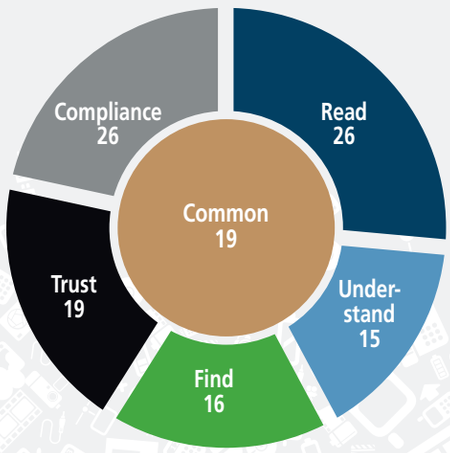
A survey performed in the LongRec project investigated what cost factors where actually considered as most important. The respondents indicated that four main areas were generally considered as quite expensive: storage, technology, security, and staff.

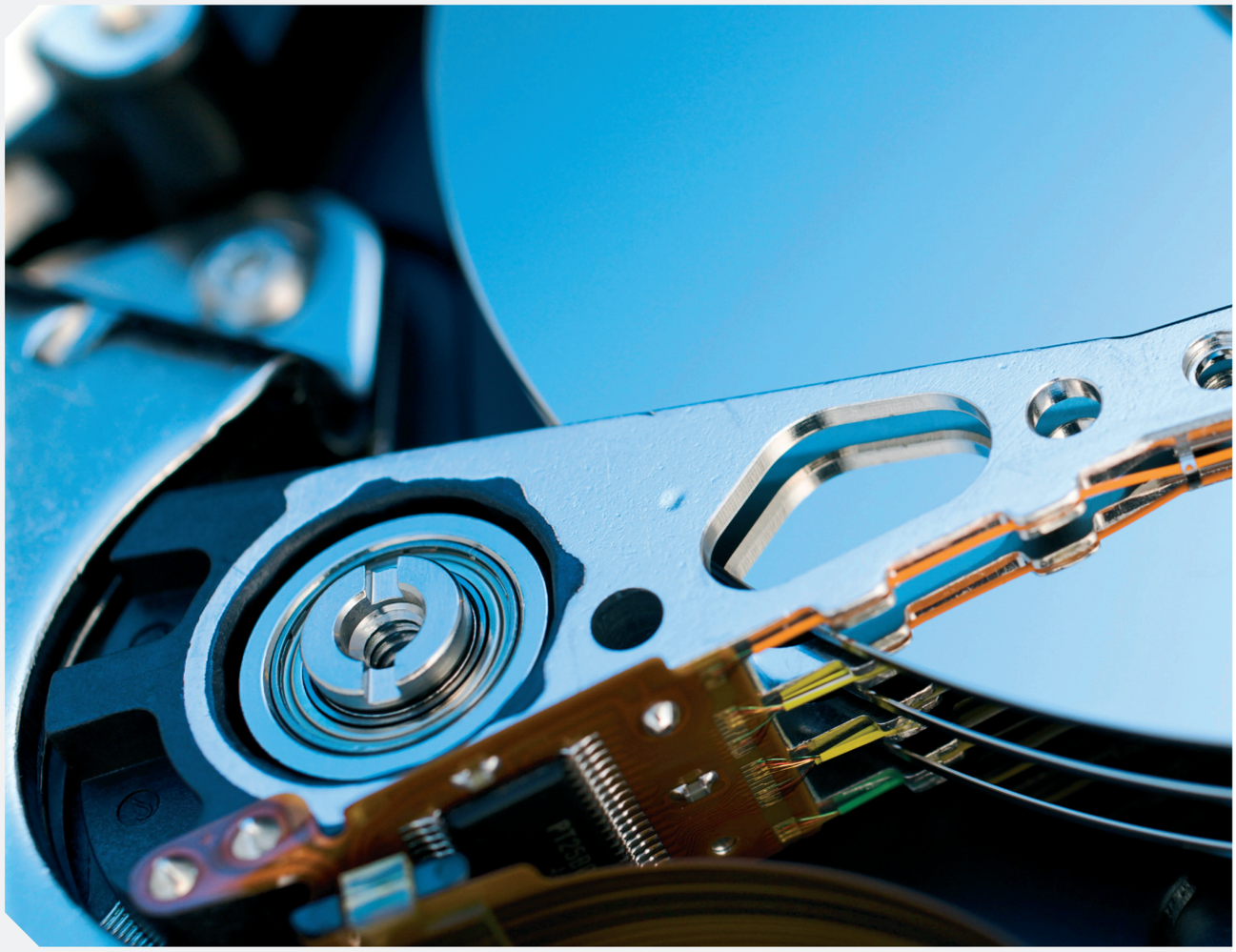
The other areas, like legal or organizational costs where not perceived as major cost driving factors *(Ref: 25).*



METADATA CHECKLIST

The successful preservation and usage of electronic records over time requires that enough and appropriate metadata are available. A comprehensive list of metadata for the READ, FIND, UNDERSTAND, TRUST and Compliance work packages was compiled providing a condensed overview of all the key aspects for various aspects in long-term data storage. *(Ref: 26)*





LIST OF THE RELEVANT STANDARDS

Dublin Core, http://dublincore.org/

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

ISO 11179, Information technology Metadata Registries

ISO 13250: 1. ISO/IEC 13250:2003 – Information technology – SGML applications – Topic maps

ISO 14001 – Environmental management systems – Requirements with guidance for use

ISO 14721:2003 – Space data and information transfer systems – Open archival information system – Reference model (OAIS) *http://public.ccsds.org/publications/archive/650x0b1.pdf*

ISO 15489 Information and Documentation – Records Management, divided in two parts: Part 1: General (ISO 15489-1:2001), Part 2: Guidelines [Technical Report] (ISO/TR 15489-2:2001)

ISO 15801:2004 Electronic imaging- Information stored electronically- Recommendations for trustworthiness and reliability

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

ISO 19005-1:2005 Document Management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)

ISO 23081 Information and Documentation – Records Management Processes - Metadata for records

ISO 5963 – Documentation – Methods for examining documents, determining their subjects, and selecting indexing terms

ISO 8000 Data Quality developed by ISO TC184/SC4

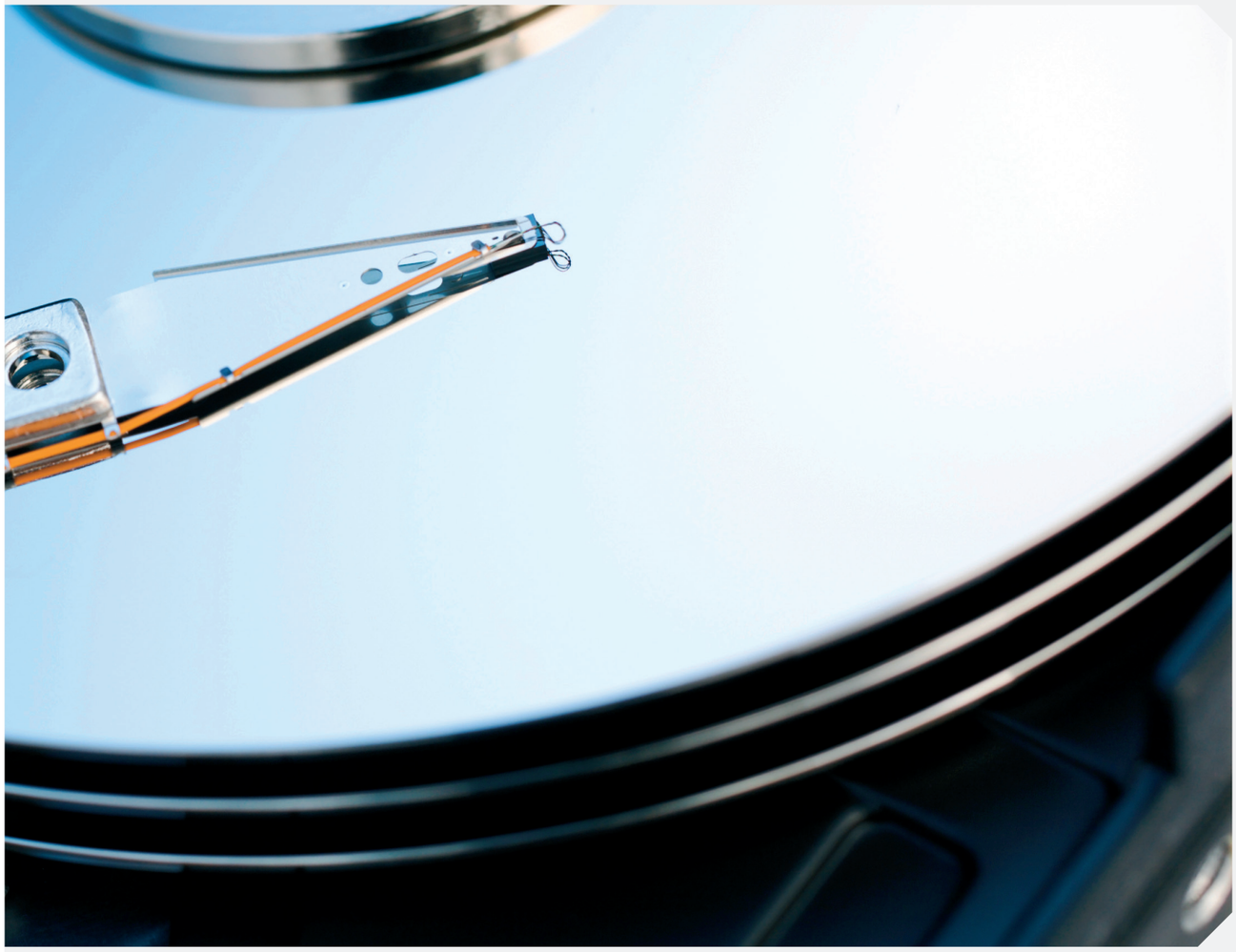
ISO 9000: 2005 Quality Management Systems – Fundamentals and vocabulary

ISO 9001- Quality management systems – Requirements

ISO/IEC 13250-2:2006 – Information technology – Topic maps – Part 2: Data model

ISO/IEC 27001:2005 (BS7799) – Security techniques – Code of practice for information security management

ISO/TR 15489-2 (provides further explanation and implementations options to achieve the outcomes of ISO 15489-1)



Moreq, Documentation on Model for Electronic Record Management

NOARK-4 *http://www.arkivverket.no/arkivverket/lover/elarkiv/noark-4.html* (5, *http://www.arkivverket.no/noark-5/utsendt_noark-5-horingsversjon_2007-2806.pdf*)

RDF – Resource Description Framework, *http://www.w3.org/RDF/*

Sarbanes Oxley Act (SOX) – some of its parts that cover rules for storage, management and search of information.

UML – Unified Modelling Language, *http://www.uml.org/*

BIP 0008-1:2004 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030104568* Code of Practice For The Legal Admissibility Of Information Stored Electronically

BIP 0009-1 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030107409* Compliance workbook

BIP 0008-2:2005 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030132417* Code of practice for legal admissibility and evidential weight of information communicated electronically

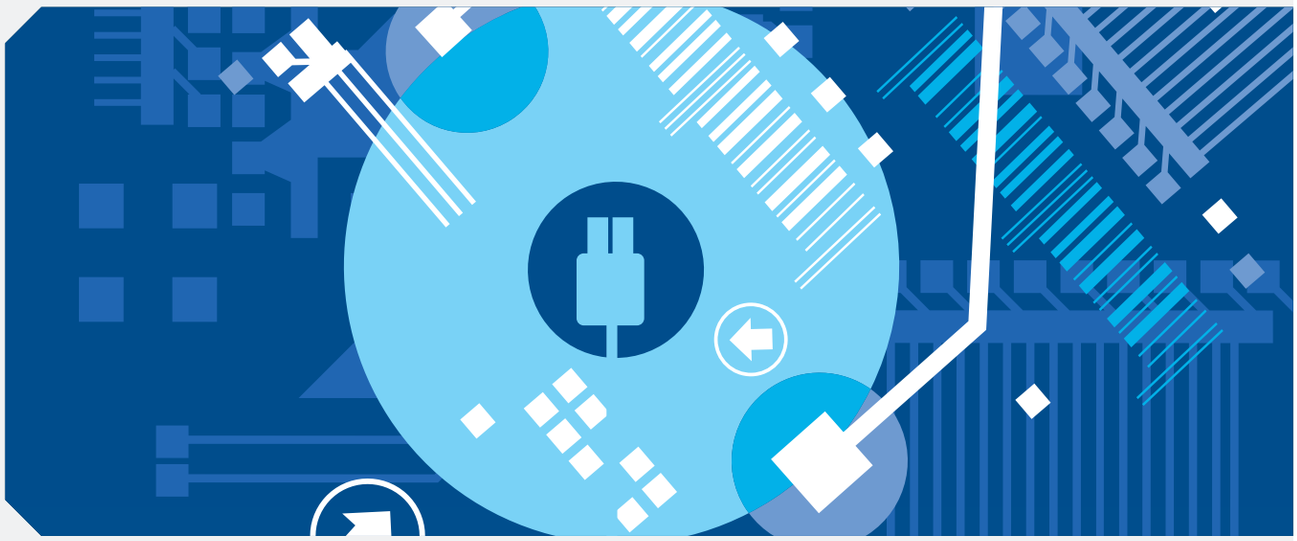
BIP 0009-2:2006 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030146667* Compliance Workbook.

BIP 0008-3:2005 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030132418* Code of practice for legal admissibility and evidential weight of linking electronic identity to documents

BIP 0009-3:2006 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030146651* Compliance Workbook.

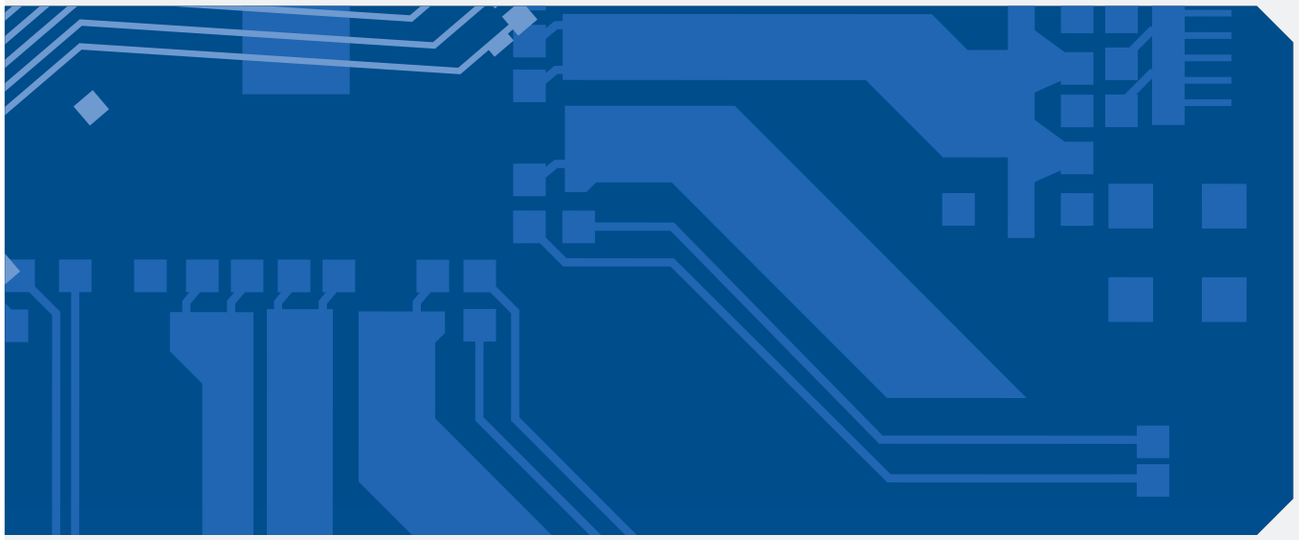
BSI DISC PD0008:1996 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000000734901* Code of Practice for Legal Admissibility of Information Stored on Electronic Document Management Systems

BSI DISC PD0008:1999 *http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030001674* Code of Practice for The Legal Admissibility and Evidential Weight of Information Stored Electronically

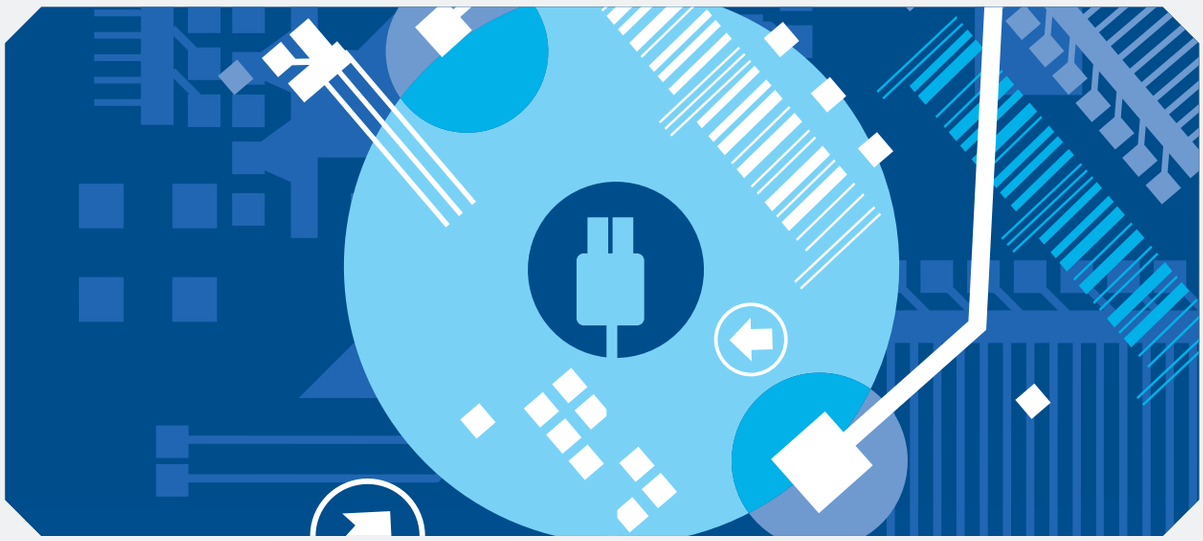


LONGREC REFERENCE LIST

1. Cerrato, O.; Gustavsen I-M.; Mestl, T.; Potterton; M., Ølnes, J.; Å lese, forstå og stole på 30 år gammel dokumentasjon – oppnåelig eller utopi? Norsk Arkivråd 3, 2007
2. Cerrato, O; Gaustad, L. (NB), Ølnes, J. Longrec Case Study: Repository Records Management. The National Library of Norway. DNV Technical Report 2008-0273
3. Mestl, T; Cerato, O; Ølnes, J; Myrseth, P; Gustavsen, I. Tid som utfordring – utfordrende tider for framtidens informasjonssøk. Norsk Arkivråd, 4/08, 2008
4. Haderlein, V; Myrseth, P; Cerrato, O. Semantic technologies in information governance, Longrec Understand DNV Research Report 2009-1950
5. Myrseth, P; Gulla, J. A; Haderlein, V; Solskinnsbakk, G; Cerrato, O. Utilizing aging public information. Scandinavian Workshop on e-Government, January 2010
6 Mestl, T; Luan, F. NB PILOT: Migration Experiments and Recommendations. DNV Technical Report: 2009-0913.
7 Luan, F; Nygård, M; Mestl, T. A mathematical framework for modeling and analyzing migration time, in Proceedings of the 10th annual Joint Conference on Digital Libraries. 2010, ACM: Gold Coast, Queensland, Australia. p. 323-332.
8. Mestl, T; Nymoen, K; Bøyum J, I; Gaustad, L. Survivability of Digital Records. DNV Tech. Rep. 2007-1623
9. Havn, W. Lotus Notes Long Time storage how to improve the ability to retain vital information for the future. HydroStatoil. 2008. (confidential)
10. Groven, A-K. Trust Strategies in Long-term Management and Preservation of Digital Records. NR-rapport number 1025, ISBN 978-82-539-0535-8
11. Luan, F; Mestl, T; Nygård, M.,"Quality Requirements of Migration Metadata in Long-term Digital Preservation Systems", accepted in 4th Metadata and Semantics Research Conference (MTSR 2010), 20-22 October 2010, Alcalá de Henares, Spain.
12. Myrseth, P; Gulla, J. A; Haderlein, V; Solskinnsbakk, G; Cerrato, O. Utilizing Ageing Information. 10th European Conference on eGovernment. June 2010.
13. Bing, J. "Mulighet for etablering av en varslingsfunksjon for bestemmer som angår dokumentadministrasjon" (note, v.2.0) Senter for rettsinformatikk, 2009
14. Mestl, T ; Gustavsen, I. M. (2008). Technology Outlook 2015: A Technology and Business Assessment within Long-Term Records Management. DNV Research & Innovation brochure.
15. Ma, J.; Abie, H.; Skramstad, T.; Nygård, M.; Requirements for Evidential Value for the Assessment of the Trustworthiness of Digital Records over Time, Proceedings of TSP2009, Macau SAR, P.R.China, Oct. 2009.
16. Ma, J.; Abie, H.; Skramstad, T.; Nygård, M.; Development and Validation of Requirements for Evidential Value for Assessing Trustworthiness of Digital Records over Time, submitted to the Journal of Theoretical and Applied Electronic Commerce Research, (Aug. 2010)
17. Ma, J.; Abie, H.; Skramstad, T.; Nygård, M.; Assessment of the Trustworthiness of Digital Records over Time, submitted to 7th ACM SAC TRECK Track, (Aug. 2010
18. Solskinnsbakk, G; Gulla, J. A; Haderlein, V; Myrseth, P; Cerrato.O. Quality of Subsumption Hierarchies in Ontologies. NLDB 2009. 14th International Conference on Applications of Natural Language to Information Systems.
19. Solskinnsbakk, G; Gulla, J. A; Haderlein, V; Myrseth, P; Cerrato.O. Quality of Hierarchies in Ontologies and Folksonomies. Accepted for publication in Data & Knowledge Engineering, 2011.
20. Gulla, J. A; Solskinnsbakk, G; Haderlein, V; Myrseth, P; Cerrato. O. Semantic Drift in Ontologies. WEBIST 2010 - International Conference on Web Information Systems.



21. Luan, F; Nygård, M; Mestl, T. "A Survey of Digital Preservation Strategies" submitted to World Digital Libraries, (Aug. 2010)
22. Gulla, J. A; Liu, J; Burkhardt, F; Zhou, J; Weiss, C; Myrseth, P; Haderlein, V; Cerrato, O. "Practical applications of Semantic Technologies", Taylor and Francis, 2010.
23. Groven, A. K. "Tillit og bevisverdi i arkivversjoner som bevares fra statlige journal/arkivsystemer" Notat no. DART/03/2010, Norsk Regnesentral, 23. mars 2010.
24. Luan, F. Migration Framework. Note, NTNU, 2010.
25. Doreen Kerubo Mageto. Cost Factors in Digital Preservation. M.Sc. Thesis in Digital Library Science at Oslo University College, 2009.
26. Mestl, T; Magnusson, J; Gustavsen, I. M; Aarnes, J; Anfindsen, O. J; Rørvik Strand; M. Gayorfar, H. Technology, Improvements and Commercialization Potential Related to Longterm Records Management. DNV Technical Report 2010-1219. (confidential)
27. Nattiya Kanhabua and Kjetil Nørvåg, Improving Temporal Language Models For Determining Time of Non-Timestamped Documents, Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries 2008 (ECDL'2008), Aarhus, Denmark, September 2008.
28. Nattiya Kanhabua and Kjetil Nørvåg, Using Temporal Language Models for Document Dating (demo paper), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'2009), Bled, Slovenia, September 2009.
29. Nattiya Kanhabua and Kjetil Nørvåg, Exploiting Time-based Synonyms in Searching Document Archives, Proceedings of JCDL'2010, Brisbane, Australia, June 2010.
30. Nattiya Kanhabua and Kjetil Nørvåg, Determining Time of Queries for Re-ranking Search Results, Proceedings of ECDL'2010, Glasgow, Scotland, UK, September 2010.
31. Nattiya Kanhabua and Kjetil Nørvåg, QUEST: Query Expansion using Synonyms over Time (demo paper), Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'2010), Barcelona, Spain, September 2010.
32. Den digitale tidsbomben tikker. VÅRT LAND, 5. april 2008 – p.28-29
33. Data for evigheita, Gemini, NTNU, 2008
34. Mestl, T; Cerato, O; Ølnes, J; Myrseth, P, Gustavsen, I. (2009) Time Challenges - Challenging Times for Future Information Search. Dlib-magazine May/June 2009 Vol. 15 No. 5/6, ISSN: 1082-9873
35. Dekeyser, H. Compliance Work Package: State-of-the-Art. ICRI - K.U.Leuven – IBBT, 2008 (confidential)
36. Felin, H; Gustavsen, I. M; Havn, W. Best practices for handling compliance issues: A maturity model to determine if your company is mature enough to handle compliance issues. AIIM conference, Philadelphia, USA, 2009.
37. Standard Norge. God IT Styring og Kontroll i norske foretak. Sluttrapport Del 1: Modell, Prosjekt NorSox, 2009
38. Standard Norge. God IT Styring og Kontroll i norske foretak. Sluttrapport Del 2: - Praktiske verktøy for innføring av God IT styring og Kontroll, Tillegg A,B og C, Prosjekt NorSox , del 2, 2009
39. Cerrato, O; Haugtomt, H (CSAM); Berge I (CSAM). Ølnes, J. CSAM LONGREC CASE STUDY: CORRECTNESS AND COMPLETENESS OF THE ELECTRONIC PATIENT RECORDS IN A PORTAL SOLUTION. DNV Technical Report 2008-1205
40. Gustavsen, I. M; Havn, W (SH); Sørdal Fosen, A. E. (SH), Cerrato, O. STATOILHYDRO LONGREC COMPLIANCE IN RETENTION AND DISPOSITION OF ELECTRONIC RECORDS (CONFIDENTIAL), DNV Technical Report 2008-1100
41. Cerrato, O. Longterm Record Management, presentation at Ark Conference, London 2007



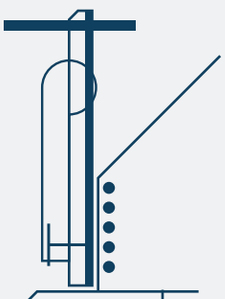
LONGREC REFERENCE LIST

42. Gustavsen, I. M. Corporate Governance, Compliance. presentation at EDOK, Conference 2007
43. Gustavsen, I. M. Informasjonshavet, IT-sjefens nye rolle, presentation at Confex Conference, 2007
44. Gustavsen, I. M. Informasjonshavet, presentation at Software Conference, 2008
45. Gustavsen, I. M. Den digitale tidsbomben, presentation at BBS days, 2008
46. Gustavsen, I. M. Den digitale tidsbomben, presentation at National Archive, 2008
47. Gustavsen, I. M. Den digitale tidsbomben, presentation at DnD frokostmøte, 2008
48. Gustavsen, I. M. Den digitale tidsbomben, presentation at Elektronisk Informasjonsforvaltning, 2008
49. Gustavsen, I. M. Den digitale tidsbomben med fokus på trender, presentation at Proact IT Norge, 2008
50. Gustavsen, I. M. The Digital Time Bomb. presentation at Arbeids og inkluderingsdepartementet, 2008
51. Gustavsen, I. M. Den digitale tidsbomben, presentation at Høyskolen i Oslo, 2008
52. Myrseth, P. Brønnøysundcaset. presentation at Semantic Days Conference, 2009
53. Myrseth, P. Forvaltning av arkivverdig data ved KS resultat XML. presentation at Sak & portaldagene, Ergogroup 2009
54. Myrseth, P. Fører elektronisk samhandling til en digital tidsbombe? presentation at Det fjerde norske arkivmøte, 2009
55. Gustavsen, I. M; Felin, H. Information maturity related to compliance. presentation at Aiim Conference, 2009
56. Gaustad, L;Tanum, J-I. Best practic migration/ conversion. presentation at EDOK Conference, 2008
57. Cerrato, O. Multimedia Digital Long term storage. presentation at BAAC Conference, Tartu, 2008
58. Gustavsen, I. M; Ølnes, J. Hvordan finne, lese, forstå og stole på digital informasjon etter 50 år?. presentation at NOKIOS 2008, Trondheim
59. IngeBorg Berg (CSAM). LongRec Awareness. presentation at Health conference in Sweden, 2008
60. Arneberg, L. Vedlikehold av lagrede BankID Signaturer i 2011 (note, 2009, confidential)

The LongRec Project

www.longrec.com

This joint-industry project focuses on the challenge of persistent, reliable, and trustworthy long-term archiving of digital records, with emphasis on availability and use of information. Problems associated with these parameters typically emerge when document lifetime exceeds 20 years. This 3-year R&D project is supported by the Norwegian Research Council.



LongRec goes beyond the "digital preservation" area addressed by libraries and (public) archives, in that information also needs to be used (retrieved, updated, and verified) and is a subject to constraints related to ownership and authorizations. All parts of a record's environment (technology, processes, organizations, roles/people, and ownership) are expected to undergo several changes during the lifetime of the record.

Long-term aspects, such as preservation (not only of availability and readability, but also of semantic value, i.e. meaning, context) and evidential value (trustworthiness) are being addressed.

Research Challenges:

The main research challenges of LongRec are establishing theory, mechanisms, and technology that may enable companies to trust the storage of digital records over a time span of several decades. The accent is particularly on 'live' records, in the sense that they continue to be updated and used over time.

In LongRec the focus is on work processes related to digital records management. This is achieved by performing case studies of some of the project partners.

The main research challenges of LongRec come under the following headings:

1. **READ:** Records transitions survival
2. **FIND:** Long-term usage
3. **UNDERSTAND:** Preservation of semantic value
4. **TRUST:** Preservation of trust and security
5. **COMPLIANCE:** Legal, social, and cultural frameworks

The LongRec project funded three Ph.D. positions within READ, FIND and TRUST at Norwegian Technical University , one post.doc. in the industry focusing on UNDERSTAND issues. They will continues to contribute with research results until end of 2011.

In addition a M.Sc. thesis was completed in collaboration with the Oslo University College within Library Science

DNV
www.dnv.com

Riksarkivet
www.riksarkivet.no

Utenriksdepartementet
www.regjeringen.no/nb/dep/ud.html?id=833

Brønnøysundregistrene
www.brreg.no

NTNU
www.ntnu.no

Katholieke Universitet, Leuven
www.law.kuleuven.ac.be/icri/

InterPares3
www.interpares.org

BBS
www.bbs.no/bbs/index.cfm

Fast Search & Transfer
www.fast.no

CSAM International
www.csam.no/csaminfo/index.htm

Nasjonalbiblioteket
www.nb.no/

NR
www.nr.no

StatoilHydro
www.statoil.no/MAR/SVG01183.nsf/fs/start

University College in Oslo
www.hio.no

Det Norske Veritas
NO-1322 Høvik, Norway
Tel: +47 67 57 99 00

www.dnv.com

LongRec
0100111 0100110

DNV-2010-V081