Semantic methods for the preservation and interpretation of information over time

LongRec, Records Management over Decades

Per Myrseth
DNV Research & Innovation
per.myrseth@dnv.com

2009-05-20
Longrec project

■ Project goal
  - Persistent, reliable and trustworthy long-term archival of digital documents, with emphasis on availability and use of documents

■ Sub goals:
  - Enable transition to digital original documents and digital work processes even for information that must be available and in use over decades
  - Explore the potential for commercial products/services in this area

When you can find, read, view and trust your data, you need to understand them, and data must be fit for purpose
Project facts

- Title: Records Management over Decades
- Home page: [www.longrec.com](http://www.longrec.com)
- Financed by the Norwegian Research Consul and partners
- Budget 30 mill NOK ++ in the period: 2007-2009
- Project leader: Inger Mette Gustavsen, [inger.mette.gustavsen@dnv.com](mailto:inger.mette.gustavsen@dnv.com)
- Project owner: Det Norske Veritas
Project partners

- InterPARES 3: [http://www.interpares.org](http://www.interpares.org)
- ICRI (Interdisciplinary Centre for Law and ICT), Katholieke Universiteit Leuven
What is an information record?

Referent:

Meta Data

Usage history

......

record

JPEG/JFIF
Adobe Photoshop
XML
WordPerfect
MPEG-2
Adobe PDF
Adobe TIFF
Macromedia swf (Flash)

......

......

Meta Data
Understand and time

- Find
- Understand
- Use

User 1

Record based on T0:
- Symbols
- Reference
- Referent
- Identifiers

User 2

Record based on T1:
- Symbols
- Reference
- Referent
- Identifiers

User 3

Record based on T2:
- Symbols
- Reference
- Referent
- Identifiers

User 4

Record based on T3:
- Symbols
- Reference
- Referent
- Identifiers

Digital repository
Record base

CHANGE
Symbols, Reference, Referent, Records, Identifiers

CHANGE
Symbols, Reference, Referent, Records, Identifiers

CHANGE
Symbols, Reference, Referent, Records, Identifiers

CHANGE
Symbols, Reference, Referent, Records, Identifiers

Time

T0  T1  T2  T3
The general problem

1. Help information users understand historic data

2. Help information providers make their information understandable over time
The Pilot: Basic Problem

- Type of data in focus: National register of Business Enterprises (national master data) at Brønnøysund Business Register (BR)

- Problem:
  - no tool support for interpreting existing enterprise data in its historic context (silent knowledge in heads of senior BR-employees)
  - no information governance policy for leveraging the implicit semantics of Business Enterprise data in the future

- Today’s situation: Semantics of Business Enterprises data are stored in
  - operational procedures
  - regulations and juridical practice
  - tools and systems for registering data in the Business Register
  - Employees head
  - ER models
  - Code tables
  - import and export formats etc.
Empowering the knowledge worker (I)

Enable the interpretation of primary data over time by relating it to and presenting it with relevant secondary data

- Primary data = data in focus, e.g.
  - Norwegian business enterprise
  - Person holding a specific role in a Norwegian enterprise

- Secondary data = relevant context data, e.g.:
  - changes in law and jurisdiction
  - changes in BR’s case practice
  - changes in language use (styreforman -> styreleder)
  (can be from both internal and external sources)
  - changes in ID used on referents, concepts, terms, records
Empowering the knowledge worker (II)

- Provide target users with a graphical tool aligning primary data and relevant context data along time.
  - Target users: junior personnel in BR’s helpdesk service team
  - External users of data

- Ensure that secondary data is
  - represented more explicitly
  - closely viewable together with the primary data in a suitable graphic interface
Success Criteria

- Ability to capture and present secondary data automatically or semi-automatically
- Junior helpdesk personnel at BR perceive the pilot as a good starting point for a useful future support tool
Deliverables related to the pilot

- Pilot software including OWL-model and technical paper
- Paper on legal aspects
  - IPR to data and sw, data quality issues, liabilities, SLA, etc
- Recommendations for an Information Governance Regime for BR ensuring the preservation of semantic value over time
“Open linked data”

Enterprise History Interpreter

Editor (Protégé)

OWL model

Merge & append ontology and instances

OWL/ RDF Source

Query & make result page

Business enterprises

Primary data

Editor (Excel)

Excel => XML

instances

Secondary data

Search & choose Business Enterprise

Client (web-browser)

Lovdata

DBpedia

Dictionaries

Google Timeline

SERES

Show result page
OWL model (working draft)
Demo: Enterprise History Interpreter

- Business enterprise data
  - Norsk Hydro ASA: 914778271
  - AKER SOLUTIONS ASA: 986529551

- People and their roles in business enterprises
  - Kjell Inge Røkke

- Next Week: Retention of data at StatoilHydro

- Demo to be published at http://longrec.brreg.no/
Research challenges

- Improve positive effect of pilot to different types of users
  - Further study of users needs
- Ontology evolution
- Merging of different open data sources (ontology alignment)
- Both human and software interpretation of semantic long term data
- Detecting and measuring semantic drift of concepts
- Access to relevant open data sources
- Strong identifiers, to link different sources of data
- Flexibility in what’s primary and what’s secondary data.
Interesting Related Examples

- Hans Rosling, Global health expert; data visionary at TED Talks
  - Simulation world health and economic development from 1800 - 2050

- Google Timeline
  - Google News Timeline is a web application that organizes search results chronologically. News and other data sources on a browsable, graphical timeline.
  - http://newstimeline.googlelabs.com/

- Wolfram|Alpha.
  - Making the world's knowledge computable. Today's Wolfram|Alpha is the first step in an ambitious, long-term project to make all systematic knowledge immediately computable by anyone.
  - http://www.wolframalpha.com/

- Google Squared
  - Google Squared returns search results in a spreadsheet format. It structures the unstructured data on web pages.
  - http://www.twine.com/item/1294x98r7-v30/what-is-google-squared-it-is-how-google-will-crush-wolfram-alpha-exclusive-video
People involved in the Pilot

- Jostein Dyrkorn, Brønnøysund Register Center
- Even Thorbergson, Dr.ing, Brønnøysund Register Center
- Alyass Muhammad, Brønnøysund Register Center
- Per Fjelde, Brønnøysund Register Center
- Tor Skjørdal, Brønnøysund Register Center
- Jon Atle Gulla, Professor, NTNU
- Geir Solskinnsbakk, PhD stud. NTNU
- Veronika Haderlein, Dr., Det Norske Veritas
- Olga Cerrato, Det Norske Veritas
- Per Myrseth, Det Norske Veritas
Technical notes

We have been using:

- Protégé 3.4 editor for OWL DL ontology
- Jena to manipulate the ontology
- Sparql for querying
- Java software developed in the project
- Similie.mit.edu Timeline ajax application
Lifespan of Norsk Hydro

Test data from the Brønnøysund Register Center

http://www.lovdata.no/cgi-win/rflfiles?doc=/ar/www/lovdata/all/id-19980717-055.html&femre=regnskap%26
**Kjell Inge Røkke**

- Bronnøysund Register Center.

Had 59 roles, roles are truncated in this view.

<table>
<thead>
<tr>
<th>Year</th>
<th>Role 1</th>
<th>Role 2</th>
<th>Role 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>Member of TOMREN VERFT AS</td>
<td>Board member of SKAARFISH GROUP AS</td>
<td>Chairman of AKER RGI HOLDING AS</td>
</tr>
<tr>
<td>1988</td>
<td>Board member of SAGVEIEN 25 AS</td>
<td>Board member of FRJONOR AS</td>
<td>Board member of AKER YARDS AS</td>
</tr>
<tr>
<td>1990</td>
<td>Member of KIR HOLDING AS</td>
<td>Board member of AVANTOR AS</td>
<td>Board member of AKER YARDS AS</td>
</tr>
<tr>
<td>1991</td>
<td>Board member of HELOX EIENDOM AS</td>
<td>Chairman of STIFTELSEN AKER STADION I</td>
<td>Board member of AKER YARDS AS</td>
</tr>
<tr>
<td>1992</td>
<td>Board member of AVANTOR AS</td>
<td>Board member of STIFTELSEN AKER STADION II</td>
<td>Board member of AKER YARDS AS</td>
</tr>
<tr>
<td>1993</td>
<td>Board member of AVANTOR AS</td>
<td>Board member of AVANTOR AS</td>
<td>Chairman of AKER RGI HOLDING AS</td>
</tr>
<tr>
<td>1994</td>
<td>Board member of AKER RGI HOLDING AS</td>
<td>Representative of STIFTELSEN AKER STADION I</td>
<td>Chairman of AKER RGI HOLDING AS</td>
</tr>
<tr>
<td>1995</td>
<td>Board member of TOMRA EIENDOM AS</td>
<td>Board member of AKER YARDS AS</td>
<td>Board member of AVANTOR AS</td>
</tr>
<tr>
<td>1996</td>
<td>Board member of TOMREN VERFT AS</td>
<td>Chair of STIFTELSEN AKER STADION II</td>
<td>Board member of TRG HOLDING AS</td>
</tr>
<tr>
<td>1997</td>
<td>Board member of AKER LANGSTEN AS</td>
<td>Board member of MIDELFART HOLDING AS</td>
<td>Board member of OY</td>
</tr>
<tr>
<td>1998</td>
<td>Board member of SKAARFISH GROUP AS</td>
<td>Board member of STADION EIENDOM AS UTLEIEBYGG</td>
<td>Chairman of KVIK</td>
</tr>
<tr>
<td>1999</td>
<td>Member of Brattvaag Skipsverft A/S</td>
<td>Board member of AKER SEAFOODS HOLDING AS</td>
<td>General manager of TRG HOLDING AS</td>
</tr>
<tr>
<td>2000</td>
<td>Member of AKER LANGSTEN AS</td>
<td>Board member of AKER SEAFOODS HOLDING AS</td>
<td>Chairman of TRG HOLDING AS</td>
</tr>
<tr>
<td>2001</td>
<td>Member of SANDAKERVEIEN AS</td>
<td>Board member of AKER SEAFOODS HOLDING AS</td>
<td>Procuration of AKER RGI HOLDING AS</td>
</tr>
<tr>
<td>2002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
www.dnv.com

The end