Linked Open Data Infrastructure for Digital Humanities in Finland

Aalto Data Day, 25.5.2018
Espoo, Finland

Eero Hyvönen
Prof., Director
Aalto University and University of Helsinki
HELDIG – Helsinki Centre for Digital Humanities
http://heldig.fi

Semantic Computing Research Group (SeCo)
http://seco.cs.aalto.fi/
Contents

• **Background:** History
• **Vision:** Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)
• **Challenges:** Data Complexity & Production
• **Solution:** Components of LODI4DH
Background: History 2001-
Semantic Web Activity at W3C Started 2001

The Semantic Web
A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities
By Tim Berners-Lee, James Hendler and Ora Lassila | May 17, 2001 | ❤ 10

Happy 20th Birthday, World Wide Web
CERN on March 13 celebrates the 20th anniversary of a proposal entitled, "Information Management: A Proposal," by Tim Berners-Lee, which would become the blueprint for the World Wide Web »
March 12, 2009
Web of Pages
WWW

Web of Data
GGG
(Giant Global Graph)
Linked Open Data Cloud 2018

LODStats 2018 (http://stats.lod2.eu/):
10000 datasets, 150 billion triples
2001

Department of Computer Science

Department information

Homepage
News and events
Research
Studies
Admission
Computing facilities
Administration
Quality manual
Contact information

Helsingin yliopisto - Tietojenkäsittelytieteen laitos

Koti Yhteystiedot Laitos lyhyesti Henkilöt Palvelut Opiskelu Tutkimus Uutiset ja tapahtumat

SEMANTIC WEB KICK-OFF IN FINLAND
- ÄLYKÄS WWW SUOMESSA

Helsingin yliopisto, Porthania, sali P3
Perjantai 2.11.2001, klo 9:00-16:30

(Hyvönen, ed., 2002)
History behind this talk

- Semantic Portals for Cultural Heritage
  - 2004 MuseumFinland – Finnish Museums on the Semantic Web
    » http://www.museosuomi.fi
  - 2008 CultureSampo – Finnish Culture on the Semantic Web 2.0
    » http://www.kulttuurisampo.fi
  - 2011 BookSampo – Fiction Literature on the Semantic Web
    » http://www.kirjasampo.fi
  - 2012 TravelSampo -- Mobile Contextualized Services of Cultural Tourism
  - 2015 WarSampo – Finnish World War II on the Semantic Web
    » http://www.sotasampo.fi
  - 2018 Semantic National Biography
    » To be published in September 2018
• Ontology and Data Services
  – 2009 National Ontology Library Service ONKI
    » [http://onki.fi](http://onki.fi)
  – 2014 ONKI.fi -> Finto.fi of the National Library
    » [http://finto.fi](http://finto.fi)
  – 2014 Linked Data Finland Data Service & Tools
    » [http://ldf.fi](http://ldf.fi)
  – 2016 Finnish Ontology Service for Historical Places and Maps
    » [http://hipla.fi](http://hipla.fi)

• Publications available at:
Joint Work During 2002-2018

- Over 30 researchers at SeCo including Eetu Mäkelä, Tomi Kauppinen, Jouni Tuominen, Kim Viljanen, Tuukka Ruotsalo, Osma Suominen, Matias Frosterus, Suvi Kettula, Kaisa Hypen, Erkki Heino, Petri Leskinen, Minna Tamper, Esko Ikkala, Mikko Koho, …
- Over 50 organizations involved
Developing the Digital World Together

HELDIG

Helsinki Centre for Digital Humanities

http://heldig.fi
What Is Digital Humanities?

Computer Science

Digital Humanities

Humanities

Social Sciences
Example Application:

WarSampo – Linked Death
Finnish WW2 on the Semantic Web

(Hyvönen et al., ESWC 2016)
Semantic Portal: 8 Perspectives to WW2

WarSampo
Finnish World War II on the Semantic Web

http://sotasampo.fi

Persons perspective

57 000 users in a week
Nov 22, 2017
WarSampo Linked Open Data Cloud

12M triples knowledge graph
Linked Data Infrastructure for Digital Humanities
LODI4DH

Traditional Infras: (rail)roads, electricity, …

Semantic Content Infra: Ontologies, metadata, data
Challenges: Content Complexity & Production
Problem 1: Cultural Content Complexity
- Heterogenous and Interlinked

Artifacts
Maps

Encyclopedia

Narratives

Literature

Buildings

Cultural sites

Biographies

Gallen-Kallela, Akseli (1865 - 1931)

Music

Fine arts

Fine art

Heterogenous and Interlinked Encyclopedia

Buildings

Cultural sites

Biographies

Gallen-Kallela, Akseli (1865 - 1931)

Music

Fine arts

Heterogenous and Interlinked Encyclopedia

Buildings

Cultural sites

Biographies

Gallen-Kallela, Akseli (1865 - 1931)

Music

Fine arts

Heterogenous and Interlinked Encyclopedia

Buildings

Cultural sites

Biographies

Gallen-Kallela, Akseli (1865 - 1931)

Music

Fine arts

Heterogenous and Interlinked Encyclopedia

Buildings

Cultural sites

Biographies

Gallen-Kallela, Akseli (1865 - 1931)

Music

Fine arts
Problem 2: Cultural Content Production System
- Distributed and Independent

- Museums
- Land survey
- Archives
- Linked Data
- Web 2.0 sites
- Media
- Libraries
- Citizens
Solution:
Sampo Model
"Sampo" Model for Semantic Interoperability

Content Providers

Linked Data

Semantic Metadata

Linked Data Infrastructure

Land survey

Museums

Web 2.0 sites

Archives

Media

Libraries

Citizens
How Does This Work in Practise?
Biographical Registries Collect Data about Persons

<table>
<thead>
<tr>
<th>henkilö</th>
<th>nimi</th>
<th>ammatti</th>
<th>syntymapaikka</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Akseli Gallen-Kallela</td>
<td>taiteilija</td>
<td>Lemu</td>
<td></td>
</tr>
<tr>
<td>H2</td>
<td>Gustaf Mannerheim</td>
<td>marsalkka</td>
<td>Askainen</td>
<td></td>
</tr>
</tbody>
</table>

**Person** | **Name** | **Profession** | **Birth Place**

- **H1**
  - Name: "Akseli Gallen-Kallela"
  - Profession: Artist
  - Birth Place: Lemu

- **H2**
  - Name: "Gustaf Mannerheim"
  - Profession: Marshall
  - Birth Place: Askainen

Biography Center
Art Museum Catalogs Paintings

<table>
<thead>
<tr>
<th>teos</th>
<th>nimi</th>
<th>tekijä</th>
<th>aika</th>
<th>aihe</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Mannerheimin muotokuva</td>
<td>Akseli Gallen-Kallela</td>
<td>1929</td>
<td>Gustaf Mannerheim</td>
</tr>
<tr>
<td>T2</td>
<td>Aino-triptyikki</td>
<td>Akseli Gallen-Kallela</td>
<td>1891</td>
<td>Aino, Kalevala</td>
</tr>
</tbody>
</table>

Art Museum Collection
Land Survey Organizations Know Places

<table>
<thead>
<tr>
<th>kunta</th>
<th>lääni</th>
</tr>
</thead>
<tbody>
<tr>
<td>Askainen</td>
<td>Varsinais-Suomen lääni</td>
</tr>
<tr>
<td>Helsinki</td>
<td>Uudenmaan lääni</td>
</tr>
<tr>
<td>Lemu</td>
<td>Varsinais-Suomen lääni</td>
</tr>
<tr>
<td>Turku</td>
<td>Varsinais-Suomen lääni</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Land Survey
Ontologies are Developed by Semantic Web Researchers

KOKO-ontology
Subclass Hierarchy

FinnONTO
RDF Connects and Harmonizes Linked Data into a GGG

Portal Triplestore

Serendipity: 1+1 > 2
$1 + 1 > 2$
In Principle a Piece of Cake but …

How to align concepts (URIs) used by different organizations?

How to align metadata models used by different organization?

SHARED INFRA NEEDED!
Major Components Needed
Key Components of a Semantic Web Infrastructure

1. Domain Ontologies
2. Metadata Models
3. Datasets
Component 1: Domain Ontologies
Concepts (IDs) and Data Model Glue Data Together

Solution: Shared Ontology Infrastructure
Ontology of Historical Persons
Resolving Identities

URL: http://dbpedia.org/resource/Pyotr_Ilyich_Tchaikovsky

Pyotr Tchaikovski (fi)
Pётр Ильич Чайковский (ru)
Pyotr Ilyich Tchaikovsky (en)
Pjotr Tjajkovskij (sv)
Pjotr Tsjajkovskij (no)
Pjotr Iljitsch Tschaikowski (de)
Piotr IIlitch Tchaïkovski (fr)
Piotr IIlich Chaikovski (es)
Pëtr II'ič Čajkovskij (it)
Pjotr Iljitsj Tsjajkovski (nl)
Piotr IIlitch Tchaikovsky (pt)
Piotr Czajkowski (pl)
Piotr Ilici Ceaikovski (ro)
Pjotr Iljics Csajkovszkij (hu)
Ontology of Historical Places and Maps
Hipla.fi: Searching for Place "Viipu..."
Ontology of Historical Events

"Semantic glue" of everything
Some Works of SeCo Group

- WW1 history ontology
- Finnish WW2 history in WarSampo
- Finnish History Ontology HISTO
  - Based on the Agricola Network timeline
  - Work continues …
Example of an Event on HISTO Ontology

**historiallinen tapahtuma: YYA-sopimus raukses**

<table>
<thead>
<tr>
<th>field</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>nimi</td>
<td>(fi) YYA-sopimus raukesi</td>
</tr>
<tr>
<td>tapahtumatyyppi</td>
<td>kahdenkeskiset kansainväliset sopimukset</td>
</tr>
<tr>
<td>kuuluu teemaan</td>
<td>poliittinen historia</td>
</tr>
<tr>
<td>kuka</td>
<td>Aho, Esko, Burbulis, Gennadi Edwardovitš (1945-), Suomi (1917-), Venäjä (1990-)</td>
</tr>
<tr>
<td>toiminta</td>
<td>allekirjoittaminen, irtisanominen, poliittinen päättöksenteko</td>
</tr>
<tr>
<td>tulos</td>
<td>Suomen tasavallan ja Venäjän federaation välinen sopimus suhteiden perusteista 63/1992</td>
</tr>
<tr>
<td>tapahtuma-aika</td>
<td>1992-01-20</td>
</tr>
<tr>
<td>tapahtumapaikka</td>
<td>Helsinki(1966-2008)</td>
</tr>
<tr>
<td>asiasana</td>
<td>diplomia, kansainväliset sopimukset, kansainväliset suhteet, ulkopolitiikka, YYA-sopimus</td>
</tr>
<tr>
<td>liitty</td>
<td>YYA-sopimus solmittiin Neuvostoliiton kanssa, YYA-sopimusta jatkettiin kahdellakymmenellä vuodella</td>
</tr>
<tr>
<td>syy</td>
<td>Neuvostoliitto lakkautetaan</td>
</tr>
<tr>
<td>tyyppi</td>
<td>historiallinen tapahtuma</td>
</tr>
</tbody>
</table>
Ontology of Historical Keywords
Aligning ontologies:
General upper ontology YSO + domain-specific ontologies

<table>
<thead>
<tr>
<th>Intersecting ontologies</th>
<th>Number of equivalent concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>YSO + TAO</td>
<td>1990</td>
</tr>
<tr>
<td>YSO + MAO</td>
<td>2330</td>
</tr>
<tr>
<td>YSO + VALO</td>
<td>950</td>
</tr>
<tr>
<td>MAO + TAO</td>
<td>1190</td>
</tr>
</tbody>
</table>

[Hyvönen et al., ESWC 2009]

http://finto.fi
Examples

- Professions
- Archeological concepts
- Object types, materials etc.
- Mappings with international thesauri
Ontology of Historical Times and Periods
Component 2: Metadata Models

- **Object and document centric**
  - Traditional approach in museums
    - Dublin Core
  - Event-centric
    - Object = sum of the events related to it (= biography)
    - New standards for data harmonization and enriching
      - CIDOC CRM (museums) + IFLA LRM (libraries)
  - Process-centric?
    - Next level after events
Example

Document/Object-centric model

Type: Image
Title: Allied Leaders at Yalta
Date: 1945
Publisher: United Press International (UPI)
Source: The Bettmann Archive
Copyright: Corbis
References: Churchill, Roosevelt, Stalin

Metadata

Photos, Persons

About...

[Slide by: Stephen Stead]
Event-centric view: CIDOC CRM

Explicit Events, Object Identity, Symmetry

E39 Actor

E52 Time-Span
February 1945

E53 Place
7012124

P11 participated in
E7 Activity
"Crimea Conference"

P7 took place at

P82 at some time within

E65 Creation

P86 falls within

E38 Image

P67 is referred to by

E31 Document
"Yalta Agreement"

P14 performed

P81 ongoing throughout

P94 has created

(Stephen Stead, 2008)
CIDOC CRM: Event Based

Top-level classes useful for integration

E55 Types
  refer to / refine

E39 Actors
  participate in

E28 Conceptual Objects
  affect or / refer to

E18 Physical Thing
  location

E2 Temporal Entities
  within

E52 Time-Spans
  at

E53 Places

(Stephen Stead, 2008)
Component 3: Datasets for the Finnish DH Cloud
Examples

- Persons
  - Semantic National Biography
- Places
  - Nimiaristo (Kotus)
  - Map collections (National Archives, Land Survey,…)
- Events
  - Acricola and other timelines of Finnish history
- Museums and Libraries
  - Finnish National Gallery
  - Finna, Fennica
  - BookSampo, CultureSampo
- Media
  - YLE and KAVA archives
- Legislation
  - Semantic Finlex
How to Deploy the Infrastructure for the Clients to Use?
How to publish Linked Data?
5-star Linked Data model

- ★ Make data available on the Web in whatever format under an open license.
- ★★★ Make data available as structured data (e.g., Excel instead of an image scan of a table).
- ★★★★ Use non-proprietary formats (e.g., CSV instead of Excel format).
- ★★★★★ Use URIs to denote things, so that people can point at your data.
- ★★★★★★ Link your data to other data to provide context.

(Tim Berners-Lee)
http://5stardata.info
Our "7-star" model and LDF.fi data hotel

Goals: enhance re-usability and data quality

7-star Linked Data Service

However, in our opinion, providing 5-star Linked Data is just the beginning. To actually make use of the datasets, consumers need more support in getting to know and access them, as well as a better grasp of their quality and provenance. To this end, we extend the model with two additional stars:

⭐⭐⭐⭐⭐ Provide your data with a schema and documentation so that people can understand and re-use your data easily.
⭐⭐⭐⭐⭐ Validate your data and denote its provenance so that people can trust the quality of your data.

This added support should come with as little extra work as possible to the data publisher. Our hypothesis is that a lot of this can be done automatically, basing on the Linked Data core. A data publisher needs only to provide their data in the RDF format, and the LDF.fi portal will do the rest automatically. See the overview paper (in ESWC 2014 Proceedings, Springer-Verlag) for some more details about the underlying ideas.
Why LDF.fi?

Living Laboratory for publishing Linked Open Data
- Same idea as in ontology services (e.g., ONKI http://onki.fi)
- But for data and schemas

Data Services for
- Linked datasets
- Schemas

Links to
- Related services
- Related applications

Learning Center
- For publishing and using Linked Data
Linked Data Finland

Living Laboratory Data Service for the Semantic Web

This site is the Living Laboratory of the Linked Data Finland research initiative, conducted by the Semantic Computing Research Group at Aalto University in collaboration with University of Helsinki and a large consortium of Finnish public organizations and companies.

Our goal is to make life easier for both publishers as well as consumers of structured data on the Web. We base our work on the Linked Data paradigm and stack of standards, which combines an expressive, semantic data model (RDF) with standardized access mechanisms (SPARQL and live HTTP URLs).

5-star Linked Data

The baseline of our work is the 5-star Linked Data model, proposed originally by Tim Berners-Lee.

- * Make data available on the Web in whatever format.
- ** Make data available as structured data (e.g., Excel instead of an image scan of a table).
- *** Use non-proprietary formats (e.g., CSV instead of Excel format).
- **** Use URLs to denote things, so that people can point at your data.
- ***** Link your data to other data to provide context.

7-star Linked Data Service

However, in our opinion, providing 5-star Linked Data is just the beginning. To actually make use of the datasets, consumers need more support in getting to know and access them, as well as a better grasp of their quality and provenance. To this end, we extend the model with two additional stars:

- ****** Provide your data with a schema and documentation so that people can understand and re-use your data easily.
- ******* Validate your data and denote its provenance so that people can trust the quality of your data.

This added support should come with as little extra work as possible to the data publisher. Our hypothesis is that a lot of this can be done automatically, basing on the Linked Data core. A data publisher needs only to provide...
Example dataset: Finnish Law as Linked Data

Semantic Finlex
Linked Data Finland

This dataset includes data regarding Finnish legislation and court decisions. The RDF data has been converted using data from the Finlex service; we call the new dataset Semantic Finlex. Special thanks to the Ministry of Justice, Eidas Publishing Ltd, and Talinnium Corp.

License
CC BY 4.0

See possible graph-specific licenses below.

Detailed Dataset Contents

**Finnish Legislation** (URI: http://idfd.fi/finlex/statutes/ed18890039)
- Finnish acts and decrees.
  - Example resource URI: http://idfd.fi/finlex/statutes/ed18890039

**Finnish Court Decisions** (URI: http://idfd.fi/finlex/citizens/)
- Decisions of the Supreme Court and the Supreme Administrative Court.
  - Example resource URI: http://idfd.fi/finlex/citizens/court6KQ

Schemas Used

Following schemas (vocabularies) are used in the datasets above:
- Scheme: http://purl.org/finlex/schema/data/
- Scheme: http://purl.org/finlex/schema/citizens/

Vocabulary Usage Analysis and Quality Issues

Following analyses tell what schemas (vocabularies) are used in the dataset graphs and how they have been used. Issues on data quality are pointed out.
Additional Services

- 5-star Linked Data Services
  - *Viewing and browsing RDF*
  - *SPARQL endpoint services (using Fuseki)*
- Documentation
- Validation
- Visualization
- Data curation
  - *Automatic annotation, RDF editing, data linking*
- Sharing policies
  - *URI minting*
  - *Licensing*
- Your data?
  - *Open service for publishing useful Linked Data*
Developers View to Linked Data: Rich Internet Applications (RIA)

Application 1

Application 2

Application N

SPARQL End Point

Linked Data Service
WWW Standard Model

Client Side (Browser)

Server Side
WarSampo: two components

Linked Data Service

http://ldf.fi

Applications

http://sotasampo.fi
Conclusions
Key Lesson Learned:
*Shared infra is the key for high quality linked data*

”Intellectuals solve problems - geniuses prevent them”
Albert Einstein
More Info – Questions?

LODI4DH
https://seco.cs.aalto.fi/projects/lodi4dh/
Semantic Web & Linked Data
http://www.w3.org/standards/semanticweb/
Sampo Model & Applications
http://seco.cs.aalto.fi/publications

In English
2012
https://www.amazon.com/Publishing-
Cultural-Heritage-Synthesis-
Technology/dp/1608459977

In Finnish
2018
https://www.gaudeamus.fi/semanttine-
web/