Using Linked Data in Digital Humanities: Shift from Data Publishing to Data Analysis and Artificial Intelligence

Prof. Eero Hyvönen
Aalto University and University of Helsinki
HELDIG – Helsinki Centre for Digital Humanities
Semantic Computing Research Group (SeCo)

National eScience Symposium 2019
Amsterdam, Nov 21, 2019
Contents

• Vision 1: A ”Sampo” model for publishing linked historical data

• Vision 2: Four generations of semantic portals

• Cases: ”Sampo” series of semantic portals
Vision 1: Sampo Model for Publishing Linked Historical Data
Web of Data: Linked Open Data Cloud
Human Knowledge on the Semantic Web

LODStats 2018 (http://stats.lod2.eu/):
10,000 datasets, 150 billion triples
Big Boys Have Entered the Game: **Knowledge Graphs**

http://schema.org

- Google Knowledge Graph
- Microsoft Satori
- IBM Watson
- eBay Products
- Facebook Graph
- ...
Problem 1: Cultural Content Complexity
- Heterogenous and Interlinked

Artifacts - Maps

Encyclopedia

Narratives - Literature

Buildings

Music

Cultural sites

Biographies

Fine arts

Gallen-Kallela, Akseli (1865 - 1931)

"Kalevala, the Epic of the Finnish People"
Problem 2: Cultural Content Production System
- Distributed and Independent

Land survey

Museums

Archives

Linked Data

Citizens

Web 2.0 sites

Media

Libraries
"Sampo" Model for Publishing Cultural Heritage Contents: Everybody Wins!

Ontology & Data Infrastructure

Content Providers
- Land survey
- Museums
- Archives
- Linked Data
- Citizens
- Libraries
- Web 2.0 sites
- Media

Semantic Metadata
Machines & Developers View: SPARQL
Useful for Digital Humanities Researchers, too!

Client Side
(Browser)

Application 1

Application 2

Application N

SPARQL End Point

Linked Data Finland Service
WWW Standard Model

Server Side
http:ldf.fi
Sampo = Mythical artifact of the Finnish Epic Kalevala that gives to its owner riches and good fortune. A metaphor of technology.

Defense of Sampo
A. Gallen-Kallela, 1896
Cultural Heritage "samos" on the Semantic Web

- **CultureSampo** – Finnish Culture on the Semantic Web (2008)
- **TravelSampo** – Mobile Contextualized Services of Cultural Tourism (2011)
- **BookSampo** – Finnish Fiction Literature on the Semantic Web (2011) [2 million users in 2018]
- **WarSampo** – Finnish WW2 on the Semantic Web (2015) [230 000 users in 2018]
  - + War Cemeteries application [2017]
  - + War Prisoners application [to be published 29.11.2019]
- **BiographySampo** – Finnish Biographies on the Semantic Web (2018) [15 000 users]
- **NameSampo** – Linked Data Workbench for Toponomastic Research (2019) [30 000 users]
- **WarVictimSampo 1914-1922** – National War History [7 000 users first day]

---

- **FindSampo** – Archaeology & Citizen Science of metal detectorists, with Finnish Heritage Agency [-2021]
- **LawSampo** – Finnish Legislation and Case Law, with Ministry of Justice of Finland (2020)
- **AcademySampo** – Historical Finnish Academic People in 1640–1899 (2020)
- **ParliamentSampo** – Parliament of Finland data, new project, Academy of Finland (2020-2022)
WarSampo - Finnish WW2 on the Semantic Web

http://sotasampo.fi/en

[Hyvönen et al., ESWC 2016]
WarSampo Infrastructure: Linked Open Data Cloud

14 million triples

[Koho et al., 2019]
Linked Open Data Cloud: Human Knowledge on the Semantic Web

Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated

Wikipedia

WarSampo (14 million triples)
• https://vimeo.com/212249404
Vision 2:
Four Generations of Semantic Portals
Four Generations of Semantic Portals for Digital Humanities

1. Printed Texts
2. Online Systems for Searching and Exploring
3. Publishing as Linked Data with Tools for DH
4. Automatic Knowledge Discovery and Artificial Intelligence

Case: BiographySampo – Biographies on the Semantic Web
Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery

Eero Hyvönen
University of Helsinki, Helsinki Centre for Digital Humanities (HELDA), Finland and
Aalto University, Department of Computer Science, Finland
E-mail: eero.hyvonen@aalto.fi

Abstract. This paper discusses a shift in focus in research on Cultural Heritage semantic portals, based on Linked Data, and envisions and proposes new directions of research. Three generations of portals are identified. Ten years ago the research focus in semantic portal development was on data harmonization, aggregation, search, and browsing (“first generation systems”). At the moment, the rise of Digital Humanities research has started to shift the focus to providing the user with integrated tools for solving research problems in interactive ways (“second generation systems”). This paper envisions and argues that the next step ahead to “third generation systems” is based on Artificial Intelligence: future portals not only provide tools for the human to solve problems but are used for framing research problems in the first place, for addressing them, and even for solving them automatically under the constraints set by the human researcher. Such systems should preferably be able to explain their reasoning, which is an important aspect in the source critical humanities research tradition. The second and third generation systems set new challenges for both computer scientists and humanities researchers.

Keywords: Digital Humanities, Linked Data, Semantic portals, Data analysis, Knowledge discovery

1. Introduction

Cultural Heritage (CH) has become a most active and forms a substantial part of DBpedia® and WikiData®. The availability of BigData has boosted the rapidly emerging new research area of Digital Humanities
BiographySampo

http://biografiyasampo.fi
BiographySampo: AI Reading Biographies for the Semantic Web

https://vimeo.com/328419960
1. GENERATION SYSTEMS

PRINTED TEXTS
National Biography of Finland
ca. 13100 bios
2. GENERATION SYSTEMS

TEXTS ONLINE
Biographies online for close reading

3. GENERATION SYSTEMS

TEXTS AS (LINKED) DATA

TOOLS FOR DATA ANALYSIS
Artificial Intelligence Reading Biographies for the Semantic Web!

National Biographies of Finland
- 13,144 biographies
- Written by 977 authors
- Interlinked with 16 data sources
- 164,000 events
- Over 120,000,000 triples (data links)

Information Extraction

Natural Language Understanding

Linked Data
Data-analytic tools: 19th century generals & admirals vs. clergy
4. GENERATION SYSTEMS

AUTOMATIC KNOWLEDGE DISCOVERY & AI
Serendipity!
The occurrence of an unplanned fortunate discovery
Publication System as an Intelligent Agent

- Paradigm Shift: Passive Tools -> Active intelligent agents
- Agents **find** ”interesting” research questions in the data
- Agents **solve** research questions
- Agents can **explain** their solutions

[Hyvönen, Semantic Web Journal, 2019]
Example: Relational Search in BiographySampo
“How are Finnish Artists Related to Italy?”

Persons = ?
Occupation = Artist
Place = Italy
Connection type = ?
Conclusions: Linked Data Makes a Difference

- **End-user’s perspective**
  - Global view to heterogeneous, distributed contents
  - Automatic content aggregation
  - Semantic search & browsing
  - Recommendation links
  - Intelligent services (knowledge discovery, personalization, visualization, …)

- **Publisher’s perspective**
  - Distributed content creation
  - Enriching each other’s contents
  - Automated link maintenance
  - Shared content publication channel
  - Reusing aggregated content in other applications
But the Lunch is not Free

- More collaboration is needed -> complicates work
- Integration of semantic portals with legacy systems
- Manual annotations are costly and may not scale up
- Automatic annotation and linking lowers data quality

Source criticism and understanding limitations of data!
More Info – Questions?

Linked Open Infra for Digital Humanities in Finland: 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/semanttinen-web/