Aalto University
School of Science

SeCo

# Semantic Web Infrastructures

## CS-E4410 Semantic Web, 17.2.2021

*Eero Hyvönen*
*Aalto University, Semantic Computing Research Group (SeCo)* http://seco.cs.aalto.fi
*University of Helsinki, HELDIG* http://heldig.fi

*eero.hyvonen@aalto.fi c*

# Learning Objectives

Understand why infrastructures are needed in applications

Learn what kind of intrastructures are available

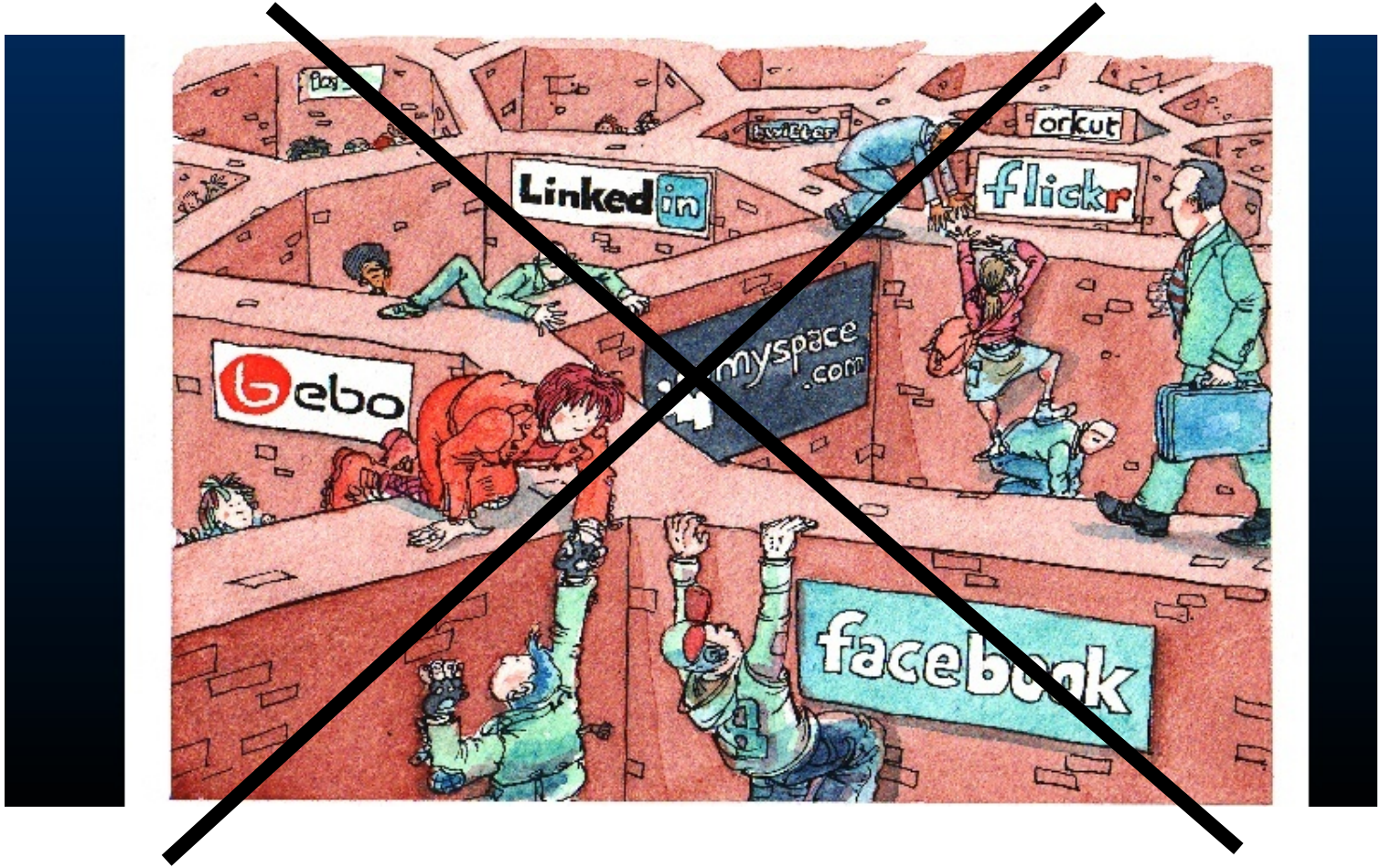
Learn about work on Finnish Semantic Web infrastructures

# Contents

- **Why infrastructures are needed?**
- **Infrastructure types**
  - Ontologies & ontology services
  - Metadata schemas
  - Data & Linked Data services
  - Software tools for developers
- **Case: Finnish Linked Open Data Infrastructure for Digital Humanities**

# Infrastructures for the Semantic Web

# Problem: Interoperability of Data

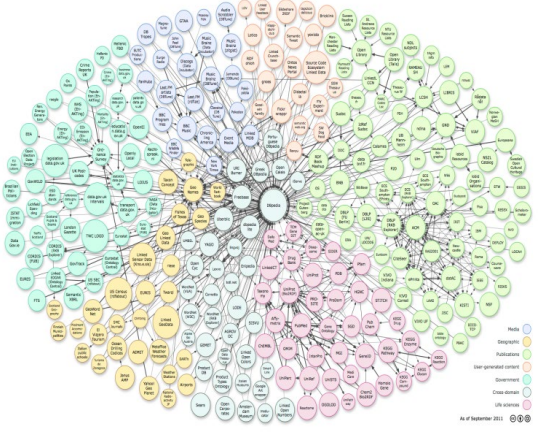


# Solution: Content Infrastructure

**Traditional Infras:**
(rail)roads, electricity, …




**Semantic Content Infra:**
Ontologies, data, metadata



# Infrastructure Types

**Shared ontologies as services**

- Creating a library service of mutually interoperable vocabularies/ontologies
- Developing the ontologies in collaboration

**Shared metadata schemas**

- Representing different information types, e.g., museum items, people, places, events

**Shared Linked Open Data & services**

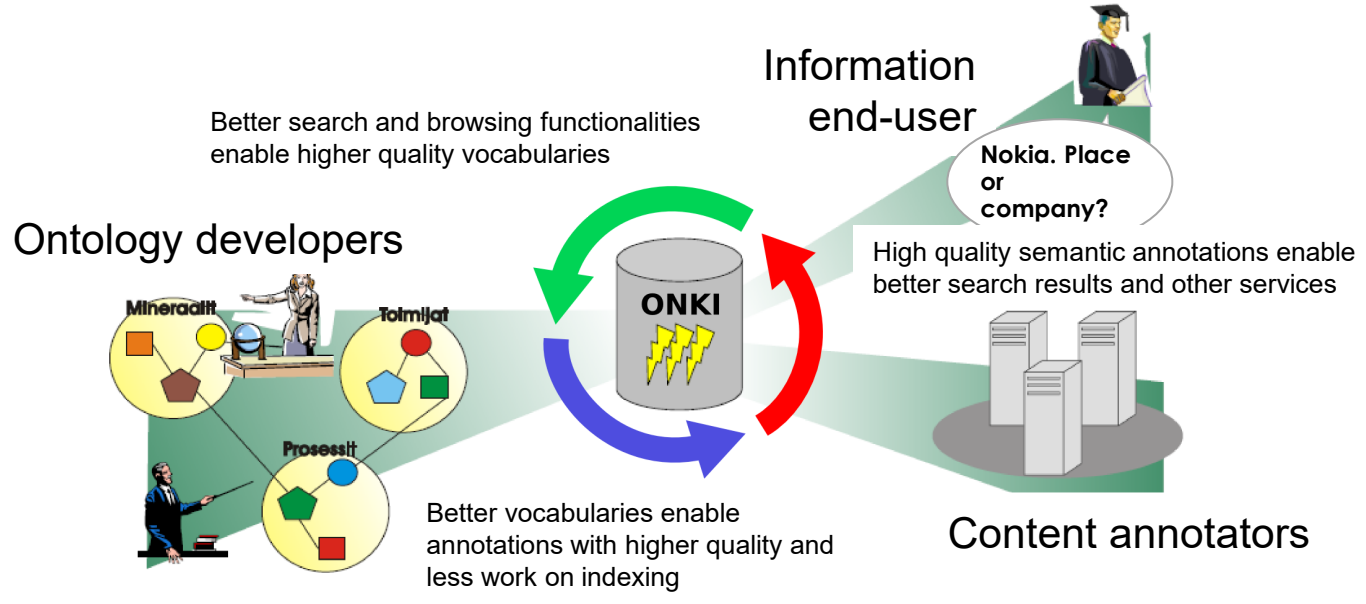
- Reducing multiple work
- Enriching each others data

**Shared software and tools**

- Re-using existing results
- Not re-inventing the wheel again and again

# Shared Ontologies & Ontology Services

# Ontology library services: ONKI concept – Users & interest groups




Supporters of the national semantic web infrastructure
**Companies**, government, **EU**, ...

# Major Domain Ontology Types

- General concept ontologies
- Actor ontologies
- Place ontologies
- Time and period ontologies
- Event ontologies
- Domain nomenclatures and terminologies
  - E.g., medical terms

- Domain ”ontology” refers thesaurus or gazetteer like KOSs whose resources are used is element values of metadata descriptions

# General Concept Ontologies

**Traditional keyword thesauri**

- General terms like ”wagon”, ”city”, ”war”, ”chair”, ...
- Correspond to classes of individuals

**Examples**

- Art and Architecture Thesaurus (AAT) (culture)
- Library of Congress Subject Headings (LCSH) (library)
- UNSPSC (products and services)
- ...

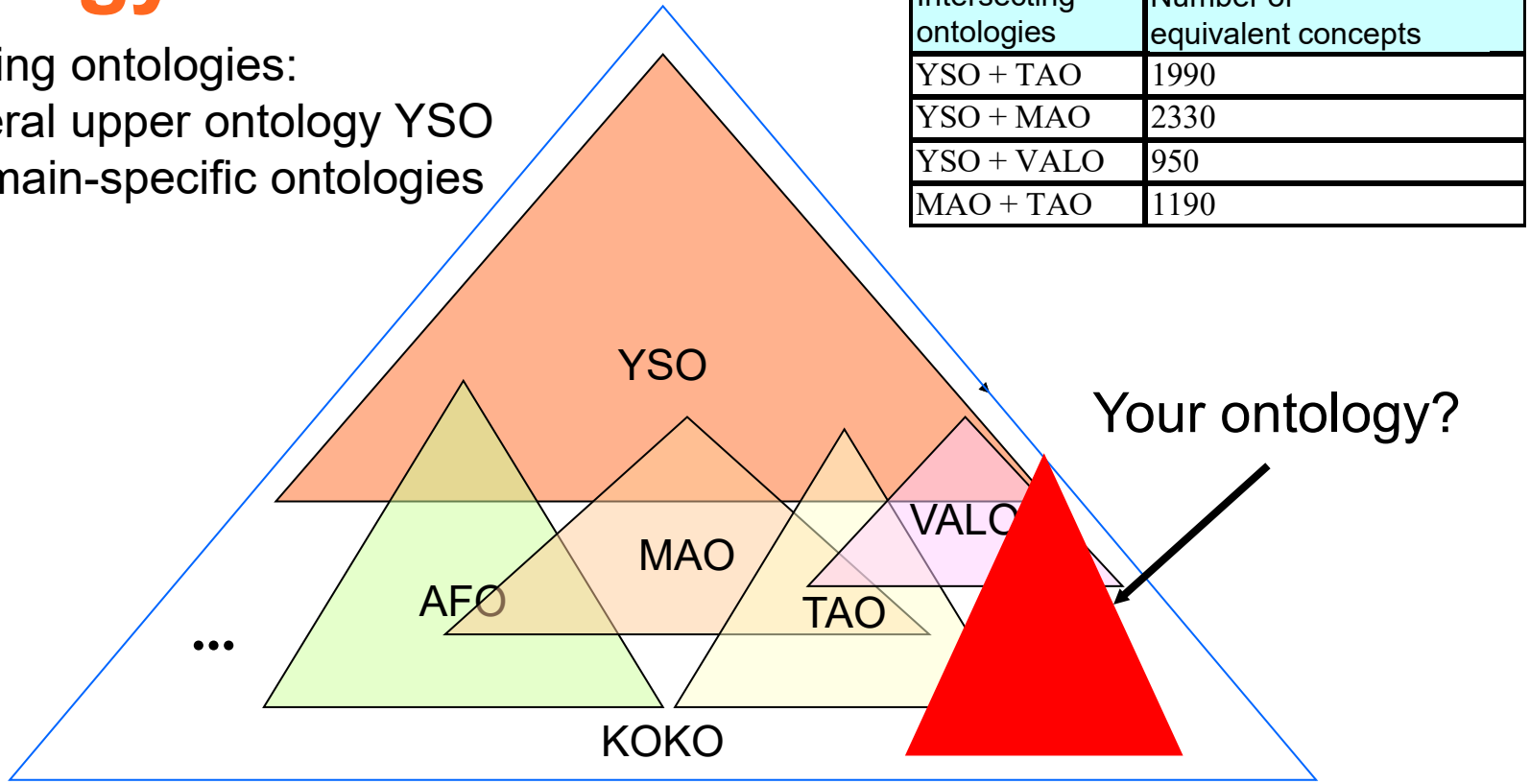
# KOKO: Linked Open Ontology cloud

| | Name | Ontology domain | Underlying thesaurus | Size | Maintaining Organization |
|---|---|---|---|---|---|
| 1 | YSO | General domain | General Finnish Thesaurus, YSA, Allärs | 23700 | National Library, Åbo Academy |
| 2 | MUSO | Music | Thesaurus of Music, MUSA/CILLA | 1000 | National Library |
| 3 | MAO | Museum domain | Thesaurus of Museum Domain, MASA | 6800 | National Board of Antiquities |
| 4 | AFO | Agriculture, foresty | Agriforest Thesaurus | 5500 | Viikki Science Library |
| 5 | TAO | Applied arts | Thesaurus of Applied Arts | 2600 | University of Eastern Finland and Library of Aalto University |
| 6 | VALO | Photography | Thesaurus of Photography Literature, Thesaurus of Photography Technology | 1900 | Finnish Museum of Photography |
| 7 | MERO | Seafaring, shipping | Thesaurus of Seafaring | 1400 | Finnish Transport Agency |
| 8 | KAUNO | Literature subjects | Thesaurus of Literature, Bella | 4900 | Finnish Public Libraries, Kirjastot.fi |
| 9 | JUHO | Public government | Thesaurus of Finnish Government, VNAS | 6400 | Ministry of Finance |
| 10 | TERO | Health promotion | YSA, TESA, MeSH, Stameta | 22000 | Various organizations |
| 11 | KITO | Literature research | Thesaurus of Literature Research | 900 | Finnish Literature Society |
| 12 | KULO | Culture research | Thesaurus for Folk Culture Studies | 1600 | Finnish Literature Society |
| 13 | KTO | Linguistics | Thesaurus of Linguistics | 1000 | Research Institute for the Languages in |
| 14 | PUHO | Defense | Thesaurus of Defence Administration | 2000 | Finnish Defence Forces |
| 15 | POIO | Points of interest | TGN, Geonames, LDG, SUO | 4600 | Various organizations |
| | **TOTAL** | | | **86300** | |

# Case: Holistic Collaborative Finnish Ontology KOKO

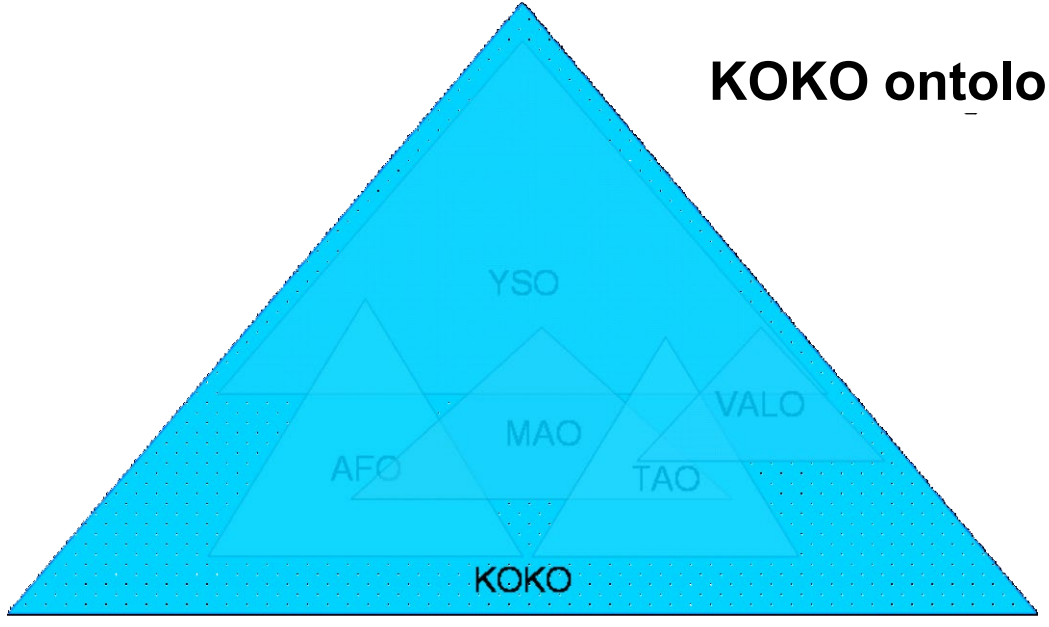
Aligning ontologies:
General upper ontology YSO
+ domain-specific ontologies

| Intersecting ontologies | Number of equivalent concepts |
|---|---|
| YSO + TAO | 1990 |
| YSO + MAO | 2330 |
| YSO + VALO | 950 |
| MAO + TAO | 1190 |




[Hyvönen et al., ESWC 2009]

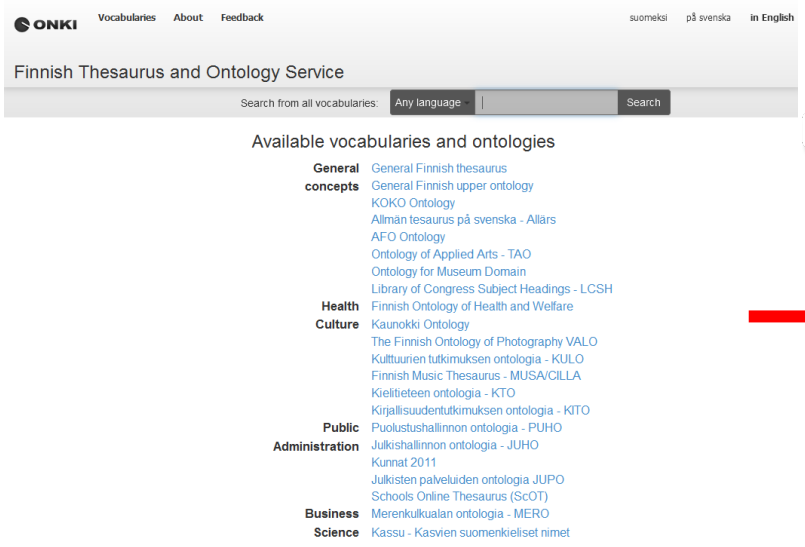
# KOKO from the “end-user” viewpoint

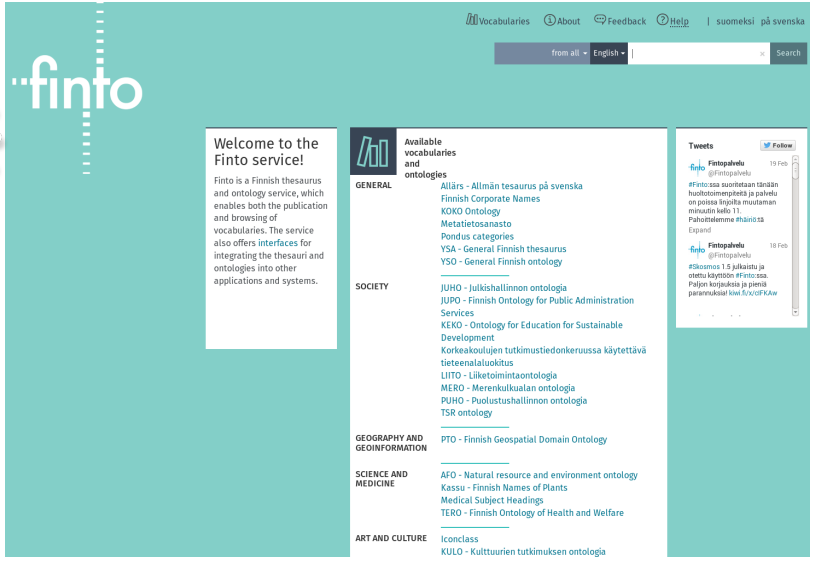


# KOKO ontologies and ONKI service deployed January 2014 by the National Library as Finto

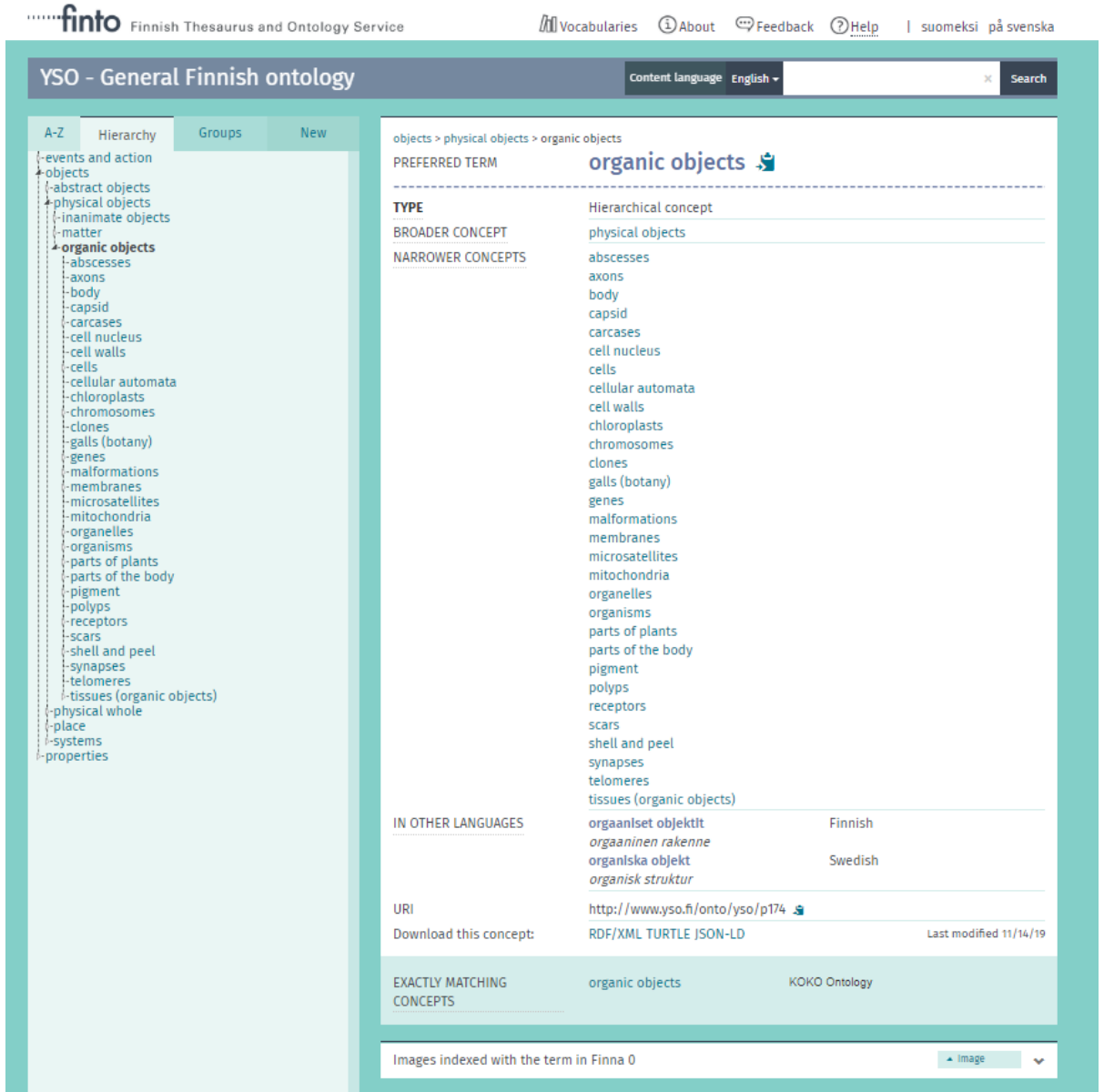
**Permanent free national service funded by Ministry of Education and Culture and Ministry of Finance**

**2019: 32 million API calls**



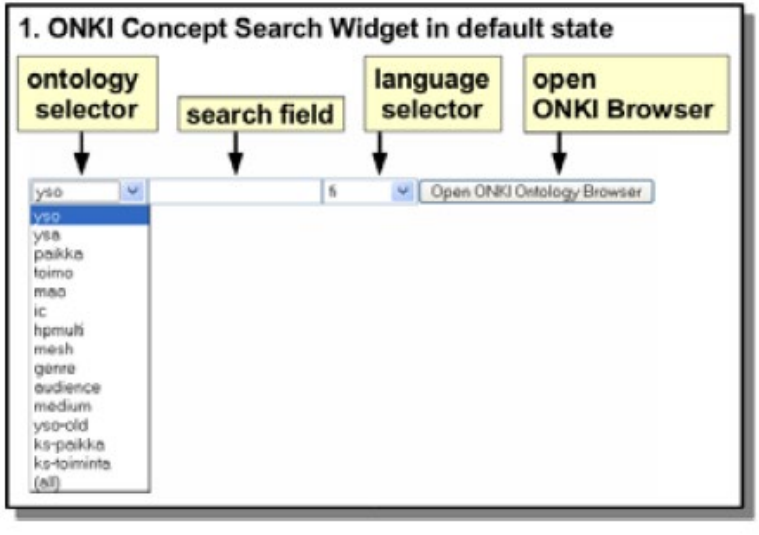
Aalto University
School of Science

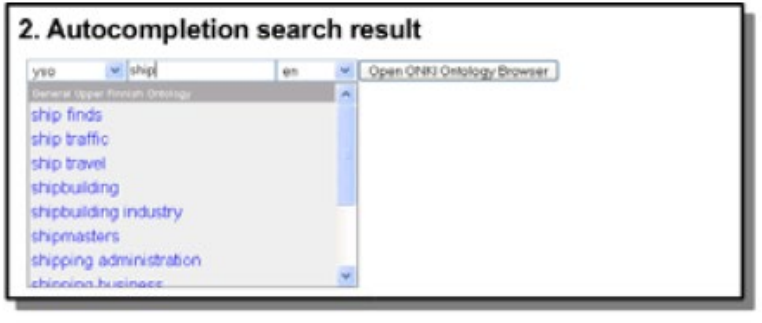


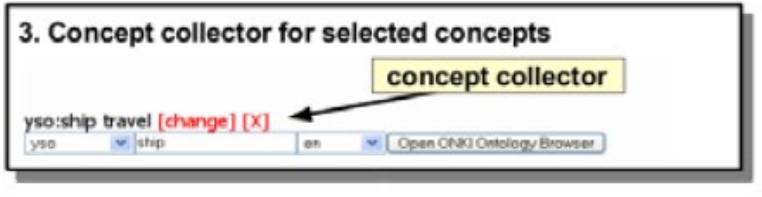

SeCo

# ONKI Widget for Mashups

- Ontology services are automatically available after publishing a vocabulary or ontology with ONKI
- Simple AJAX-based widget for creating mash-ups



# Major components of an ontology infrastructure

- Ontologies
- Ontology Library Services

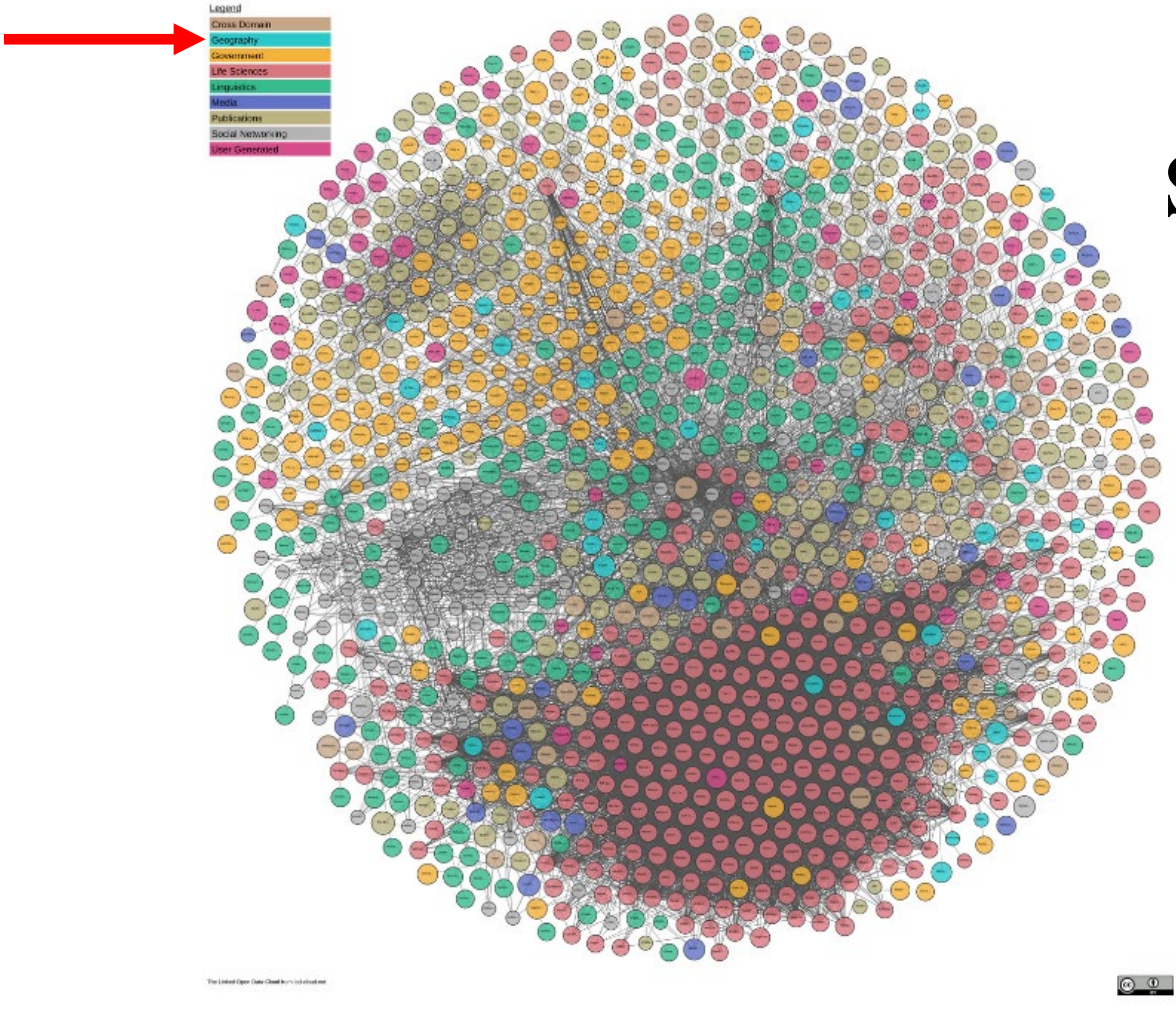
# Actor Ontologies: Resolving Identities

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

Pjotr Tšaikovski (fi)
Пётр Ильи́ч Чайко́вский (ru)
Pyotr Ilyich Tchaikovsky (en)
Pjotr Tjajkovskij (sv)
Pjotr Tsjajkovskij (no)
Pjotr Iljitsch Tschaikowski (de)
Piotr Ilitch Tchaïkovski (fr)
Piotr Ilich Chaikovski (es)
Pëtr Il'ič Čajkovskij (it)
Pjotr Iljitsj Tsjaikovski (nl)
Piotr Ilitch Tchaikovsky (pt)
Piotr Czajkowski (pl)
Piotr Ilici Ceaikovski (ro)
Pjotr Iljics Csajkovszkij (hu)

# Geography: A Key Element in the Linked Open Data Cloud

https://lod-cloud.net/

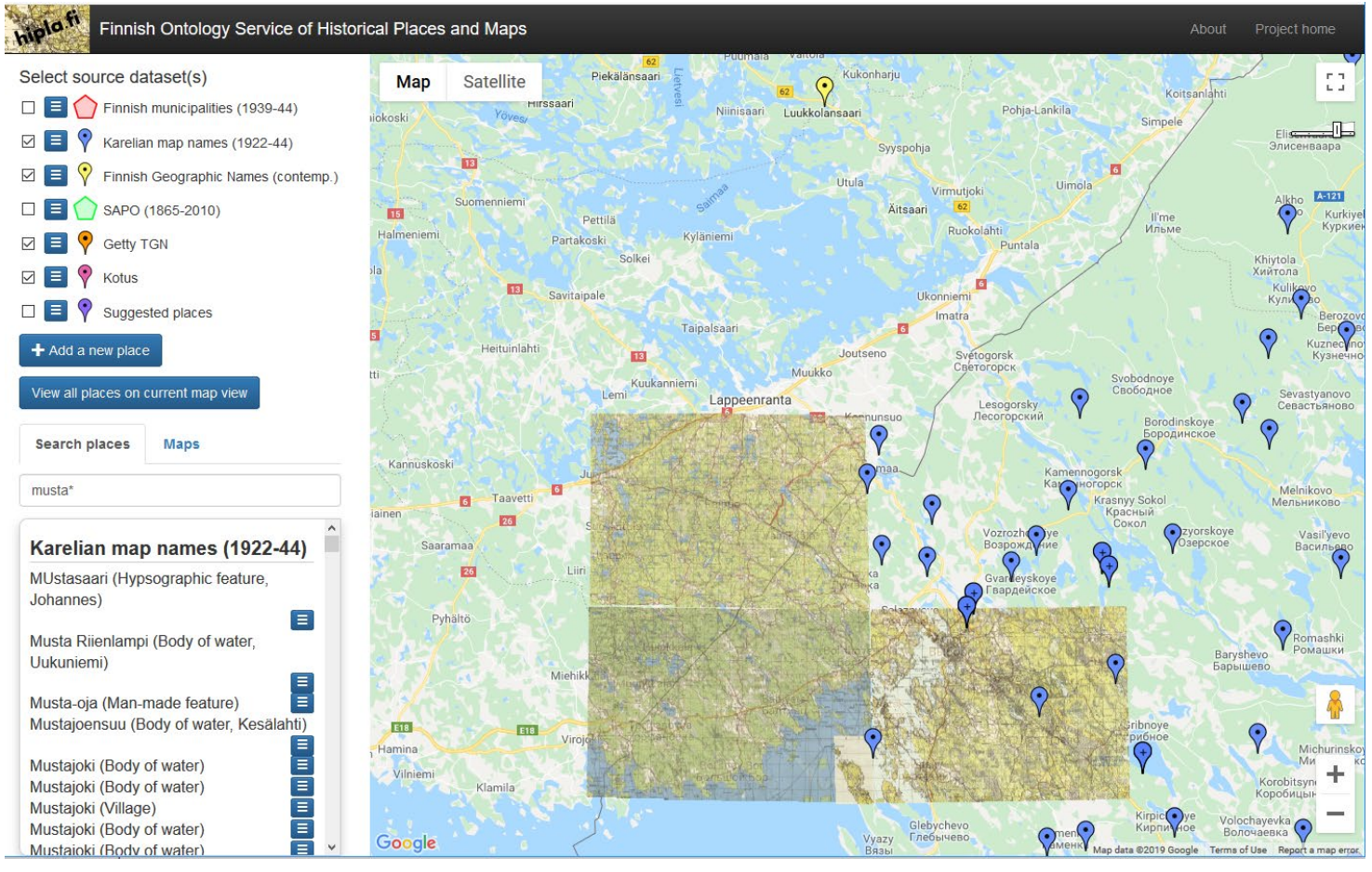



**Semantic Web**

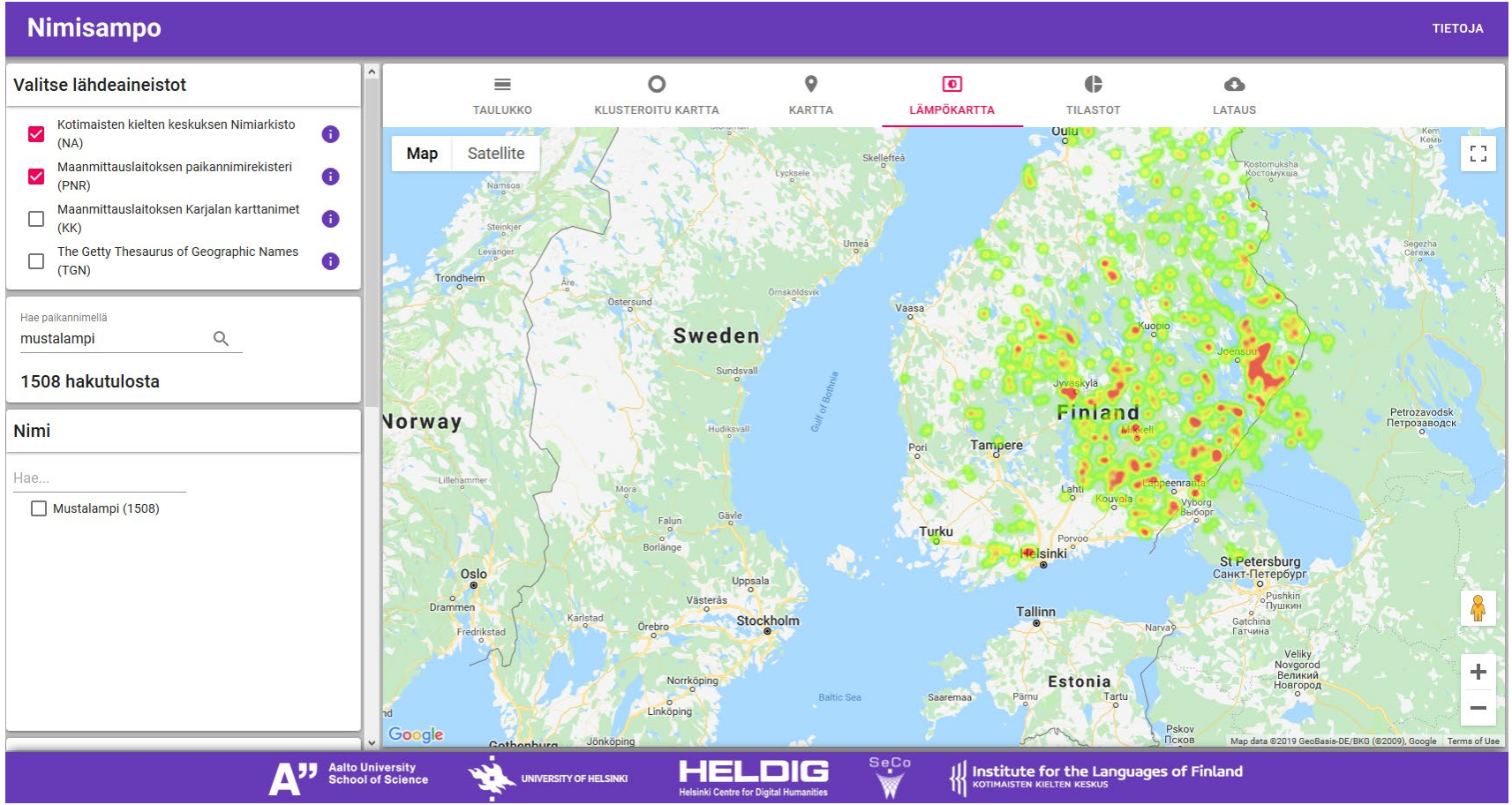
**LODstats.aksw.org:**
10 000 datasets
150 000 000 000 triples

# Finnish Ontology Service of Historical Places and Maps:

http://hipla.fi



# NameSampo: http://nimisampo.fi



# Time Ontologies

- Modeling linear and cyclic time
- Time periods are different in different countries
  - *E.g., Bronze Age in Egypt and Nordic Countries*
- Modeling uncertainty in time

# Event Ontologies

**Events are ”semantic glue” that link together:**

- **Places where events occur**
- **Times when events occur**
- **Actors who participate in events in roles**
- **Other related events**

# Shared Metadata Schemas

# Two Main Approaches

**Dublin Core approach**

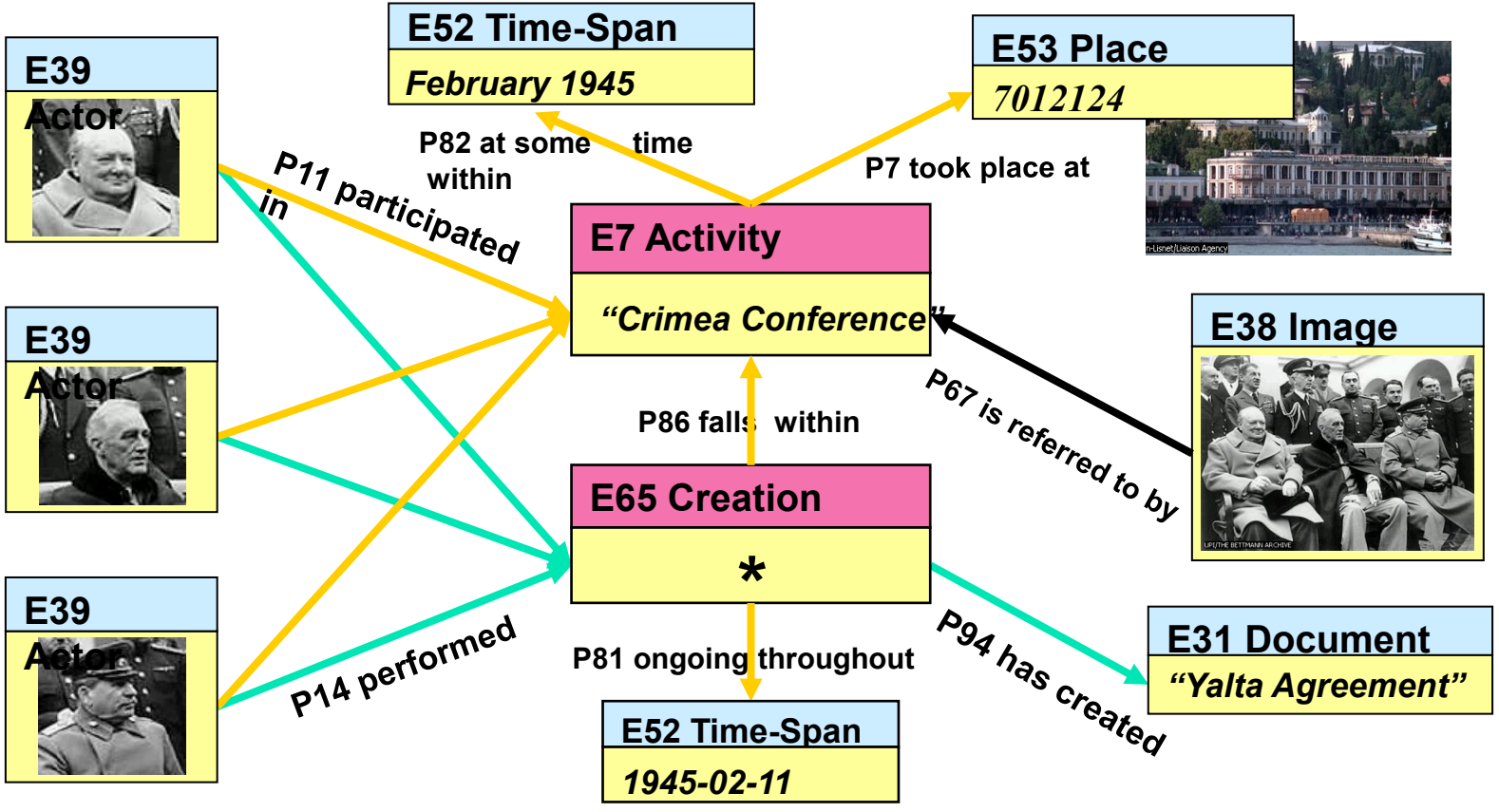
- **Mapping/refining schemas using subproperties**
- **“Dumb down principle” is used**
- **https://dublincore.org/**

**Using foundational ontology models**

- **Different schemas are mapped onto a shared ontology**
- **CIDOC CRM is a prominent standard of this**
  - http://www.cidoc-crm.org/

# CIDOC CRM:

## Using events as the foundation for knowledge representation

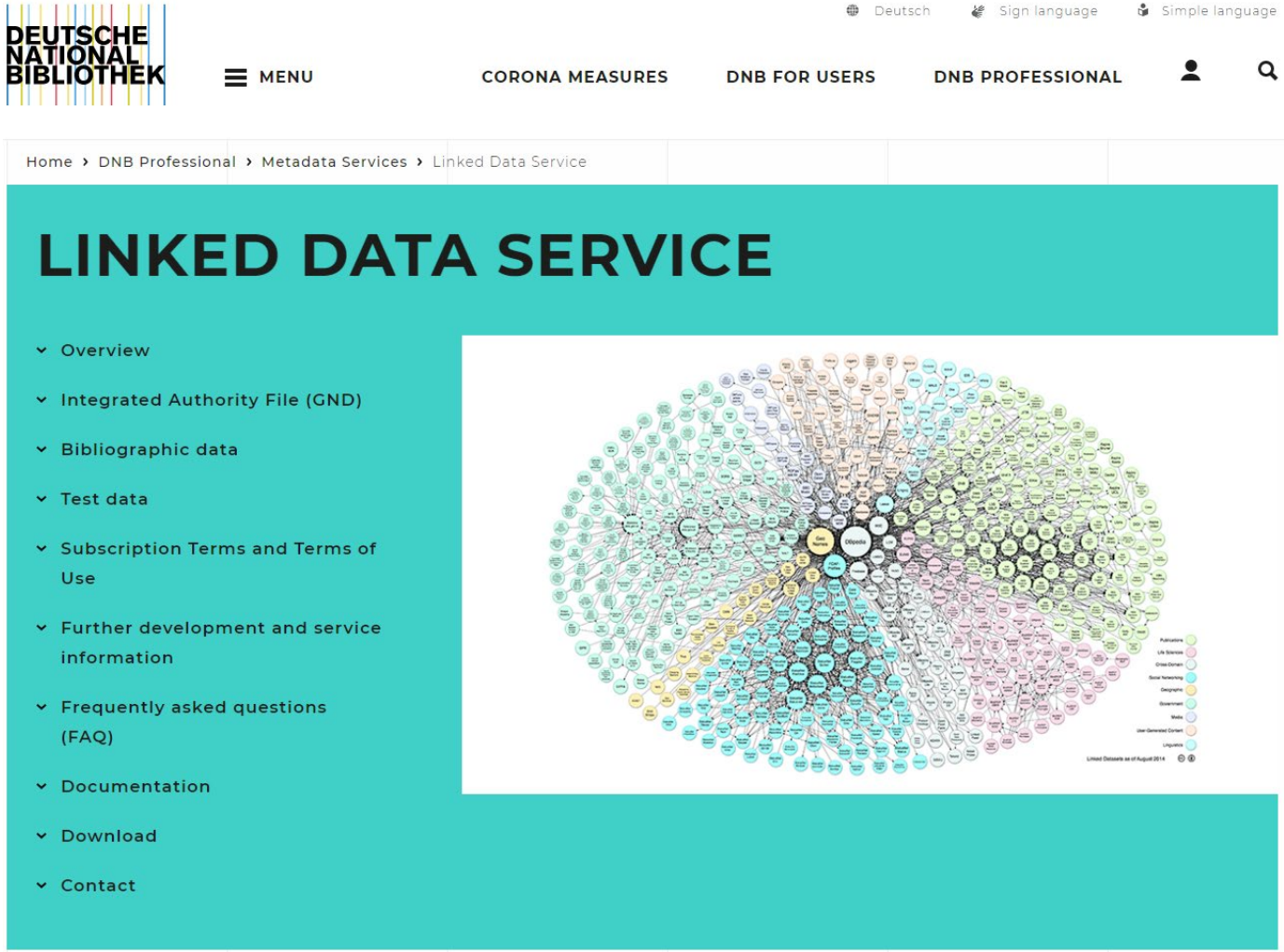



[Slide by: Stephen Stead]

# Shared Linked Open Data & Services

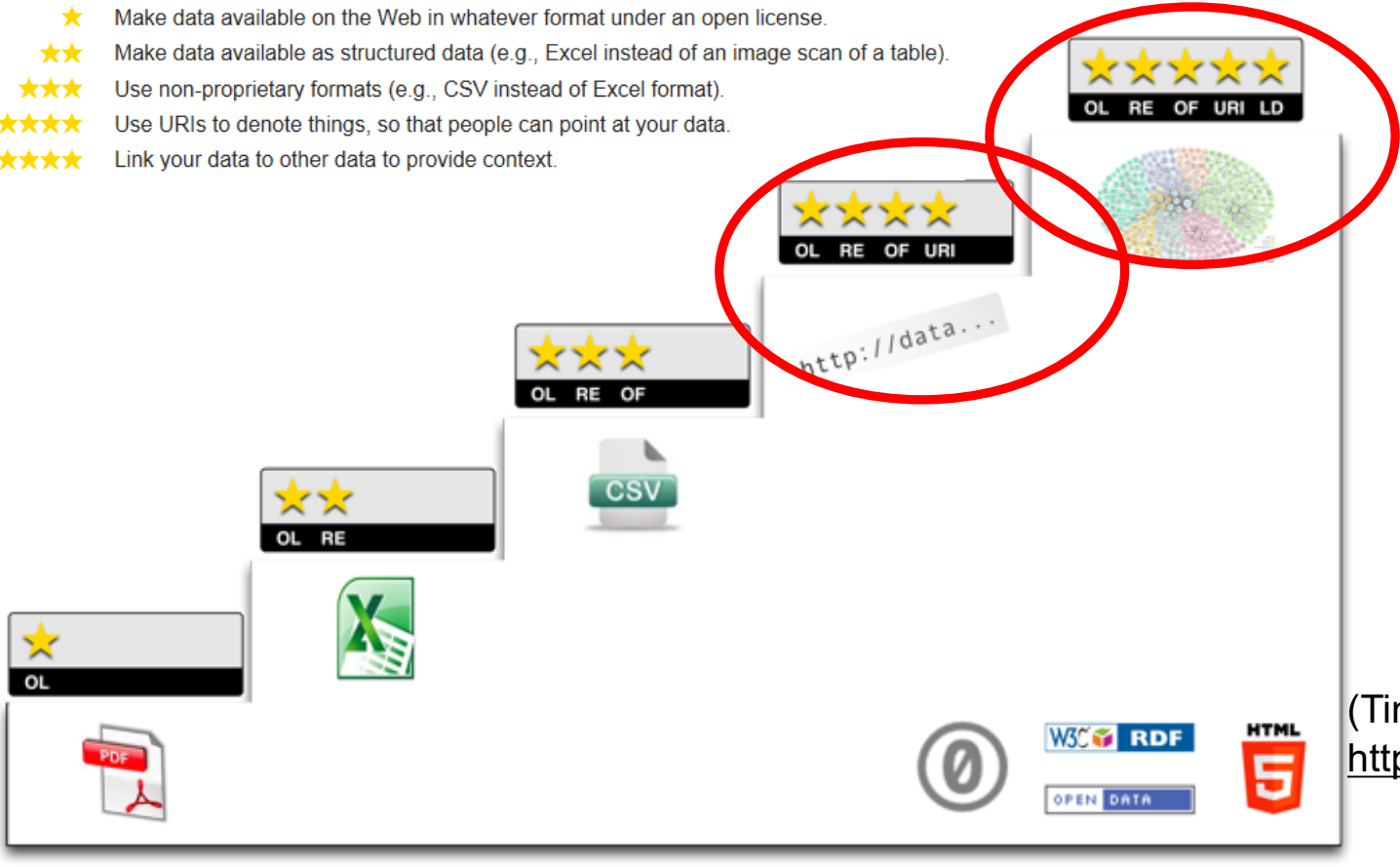
https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/LDS/lds_node.html

**An example of a Linked Data Service**



# 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

# Case: Linked Data Finland
# ”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.

Burj Al Arab

# Why LDF.fi?

## Living Laboratory for publishing Linked Open Data

- Same idea as in **ontology services**
- But for **data** and **schemas**

## Data Services for

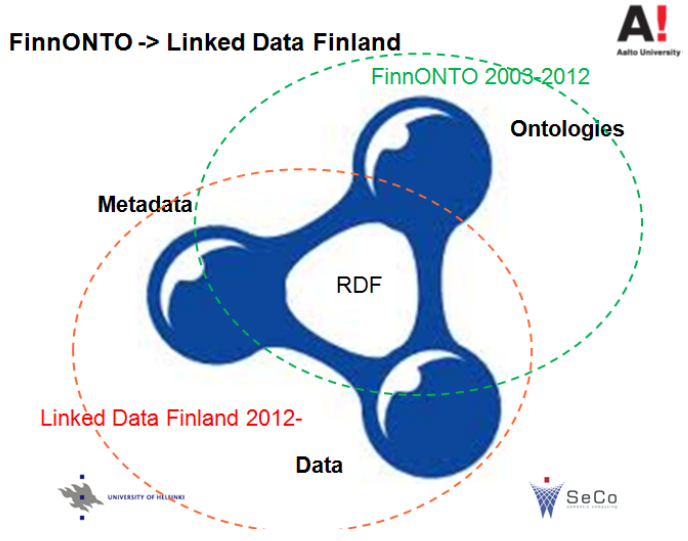
- Linked datasets
- Schemas

## Links to

- Related services
- Related applications

## Learning Center

- For publishing and using Linked Data



# Linked Data Finland Living Lab http://ldf.fi

- Home
- Project
- Datasets
- Schemas
- Services
- Policies
- Documentation
- Validation
- Applications
- Your 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 URIs).

### 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 URIs 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

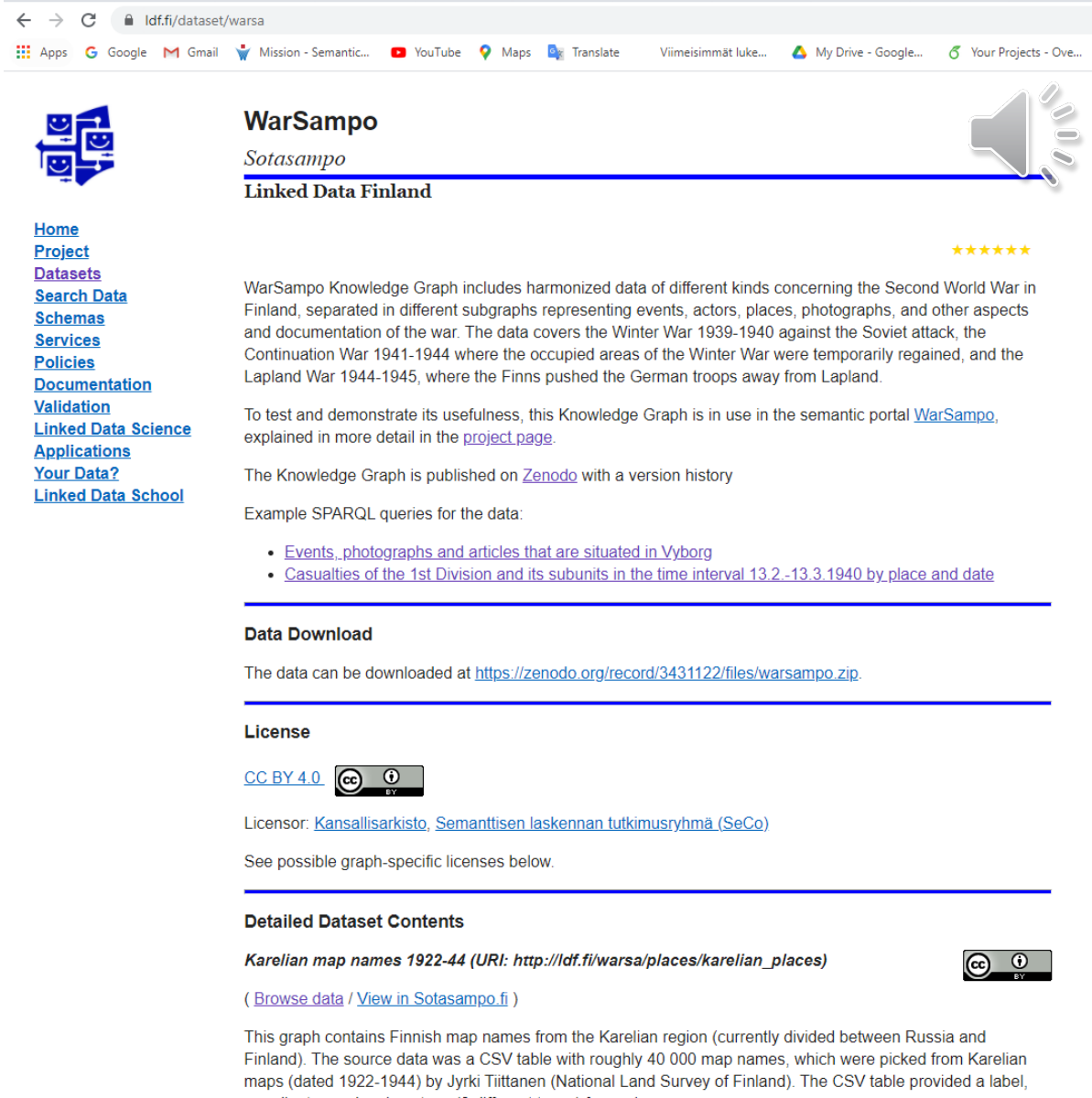
However, in our opinion, providing 5-star Linked Data is just the beginning. To actually make use of the datasets, consumers need more support in getting to know and access them, as well as a better grasp of their quality and provenance. To this end, we extend the model with two additional stars:

- ★★★★★★ 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: WarSampo Linked Data & SPARQL endpoint

**https://www.ldf.fi/dataset/warsa**



# Services

- Customary 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*

# Software Tools for the Semantic Web

Aalto University
School of Science

Department of
Computer Science

SeCo

# Component Technologies and Tools for the Semantic Web

## Languages & standards of W3C and others

- *Data exchange language:* *RDF*
- *Vocabulary/schema languages:* *SKOS, OWL*
- *Data/ontology query language:* *SPARQL*
- *Rules for reasoning:* *RIF, SWRL, …*
- *Metadata and ontology models* *DC, CIDOC CRM, …*

## Triple stores for data services

- *Fuseki, Sesame, Redland, Virtuoso, …*
- *http://en.wikipedia.org/wiki/Triplestore*

## Development tools

- Ontology editors
  - *Protégé https://protege.stanford.edu/*
  - *TopBraid Composer https://www.topquadrant.com/topbraid-composer-install/*
- Software development tools
  - *Java: Apache Jena https://jena.apache.org/*
  - *Python: RDFLib https://pypi.org/project/rdflib/*

# Case: Finnish Linked Open Data Infrastructure for Digital Humanities

# Case on Video: https://vimeo.com/460086143

# Summary

**Semantic Web infrastructures are needed**

- for data interoperability
- for reusing data, schemas, ontologies, and software

**Infrastructures include**

- Ontologies & ontology services
- Shared metadata models
- Linked data services
- Shared software and tools

**In Finland a national level solution is being developed**