Aalto University
School of Science

SeCo

Publishing Linked Data as Datasets

CS-E4410 Semantic Web, 26.1.2022

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

Learning Objective

Understand how Linked Data is published on the Web as datasets and SPARLQ endpoints to be used for practical applications

Outline

Web of Data: Basic principles

Distributed Approach: Embedding data in HTML pages

Centralized Approach: Standalone data services on servers

Web of Data: Basic Concepts

Aalto University
School of Science

Department of
Computer Science

SeCo

Basic Concepts Clarified

Linked Data

- Means practical and simple Web of Data using RDF graphs
  - Semantic Web technologies include also more advanced models
- Based on W3C's Semantic Web standards
  - *"Rebranding Semantic Web", focus on simple practical semantics*
- Linked Data is often open but can be closed, too

Open Data

- Openly available data (on the Web), under an open license
  - Cf. Creative Commons Licenses: https://creativecommons.org/
    - *"Open data and content can be freely used, modified, and shared by anyone for any purpose" – http://opendefinition.org*
- Open data is not necessarily free of charge

What is Web of Data?

The Web we see is a network that links pages: Web of Pages

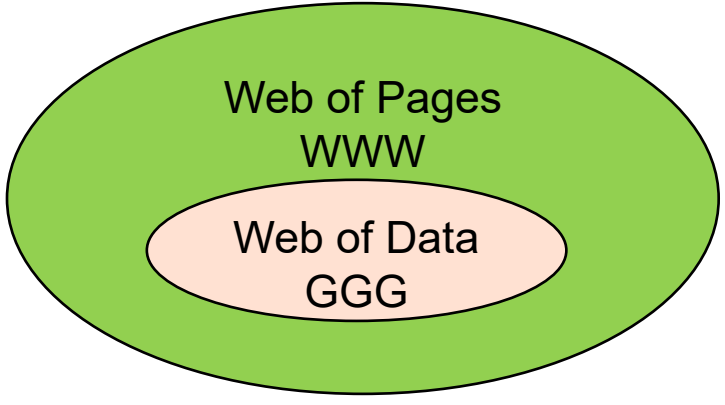
- Can be searched/browsed with a traditional web browser
- Links connect pages

A data network has emerged inside the web: Web of Data

- Can be searched/browsed with a semantic application
- Links in Web of Data connect concepts and data (e.g., eagle → bird)

Semantic Web consists of *both* networks

- Web of Pages (for humans)
  - *WWW World Wide Web*
- Web of Data (for machines to use)
  - *GGG Giant Global Graph*
  - *Examples of knowledge graphs in applications:*
    - Google Knowledge Graph, Microsoft Satori, Facebook Open Graph, …
    - Lots of domain specific applications

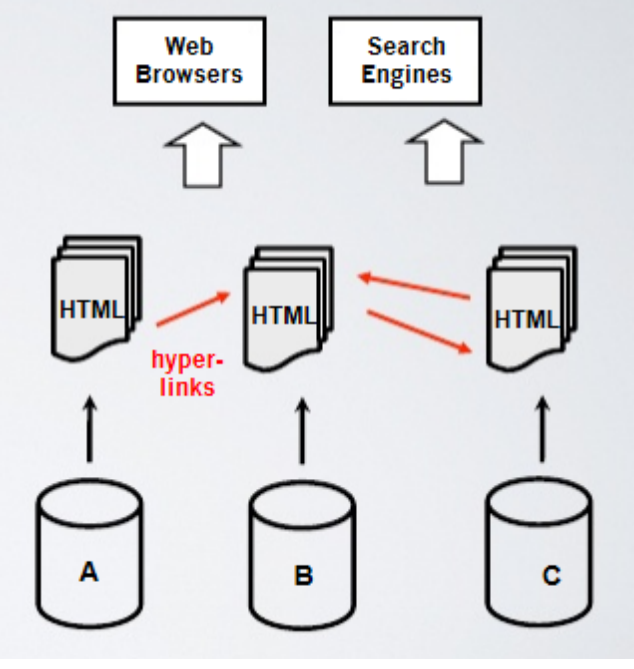


Idea of Linked Data

- Datasets are created and published using RDF
  - Embedded in web pages or as data services / data bases
  - Lightweight Semantic Web technologies are used
    - “A Little Semantics Goes a Long Way“ – Jim Hendler
- Datasets are linked together to enrich their content
  - *Cross-referencing data in other datasets*
    - E.g., place ”Finland” in GeoNames.org to president ”Niinistö” in DBpedia.org
  - *Identifying same concepts in different datasets (data alignment)*
    - E.g., ”Helsinki” in GeoNames.org vs. DBpedia.org
- Enriched data is (re-)published as data services for applications
- Coordinated by Linked Open Data (LOD) communities
  - *Application domain specific and global ones*

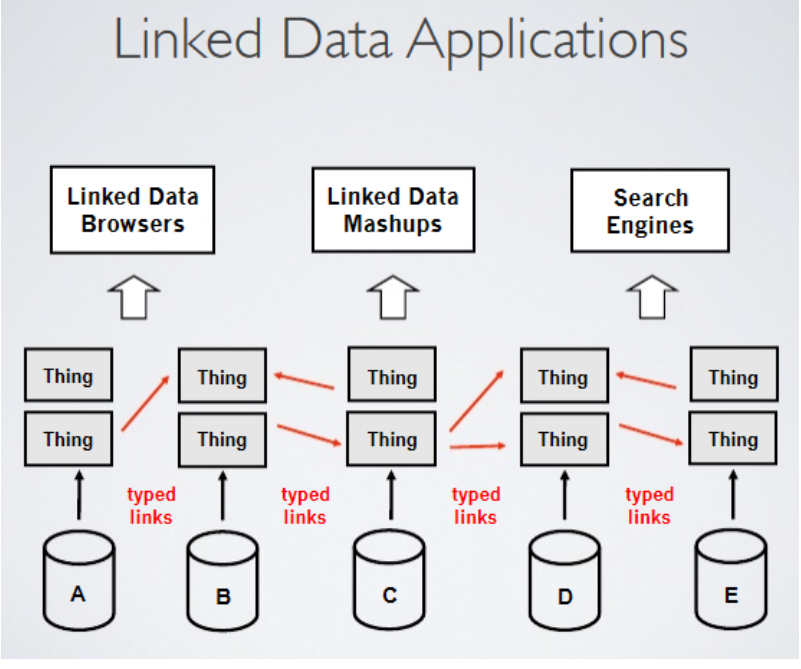
WWW and GGG coexistence

WWW



(Anja Jentzsch, 2012)

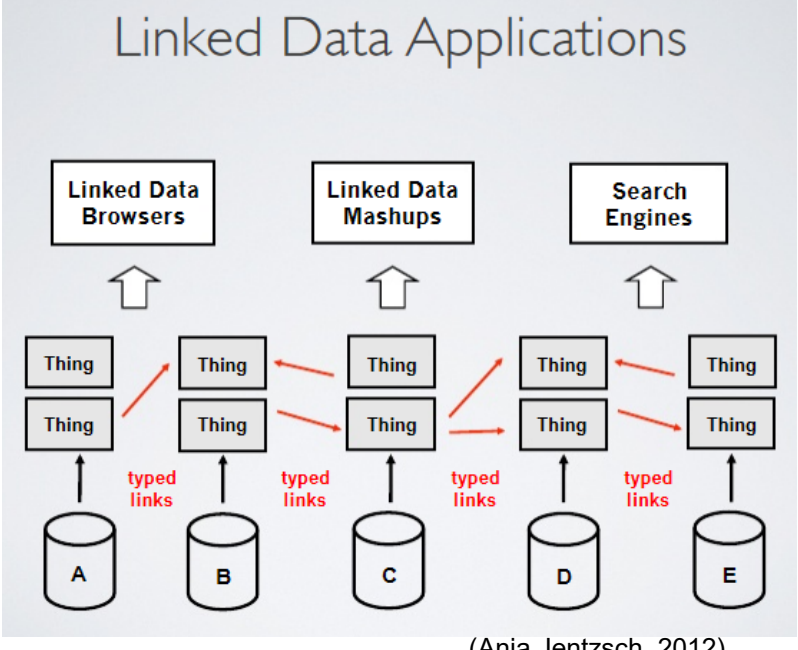
GGG



(Anja Jentzsch, 2012)

Dimensions of application development

- Services for
  - *Humans*
  - *Machines*
- Data linking
- Data aggregation
- Data harvesting
- Data production



(Anja Jentzsch, 2012)

Publishing Linked Data as Datasets for the Web of Data

Aalto University
School of Science

Department of
Computer Science

SeCo

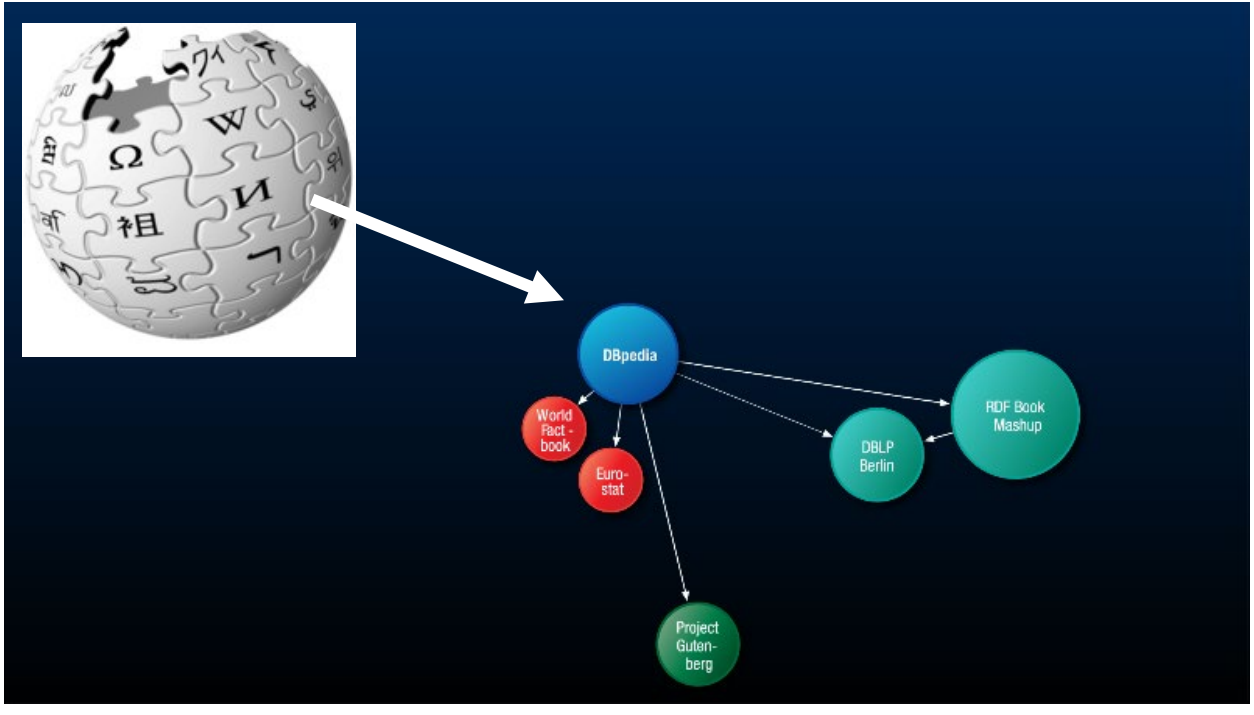
Dataset Publishing: two ways

Publish datasets for downloading or via APIs

- **CSV files**
- **RDF files**
- **JSON files**

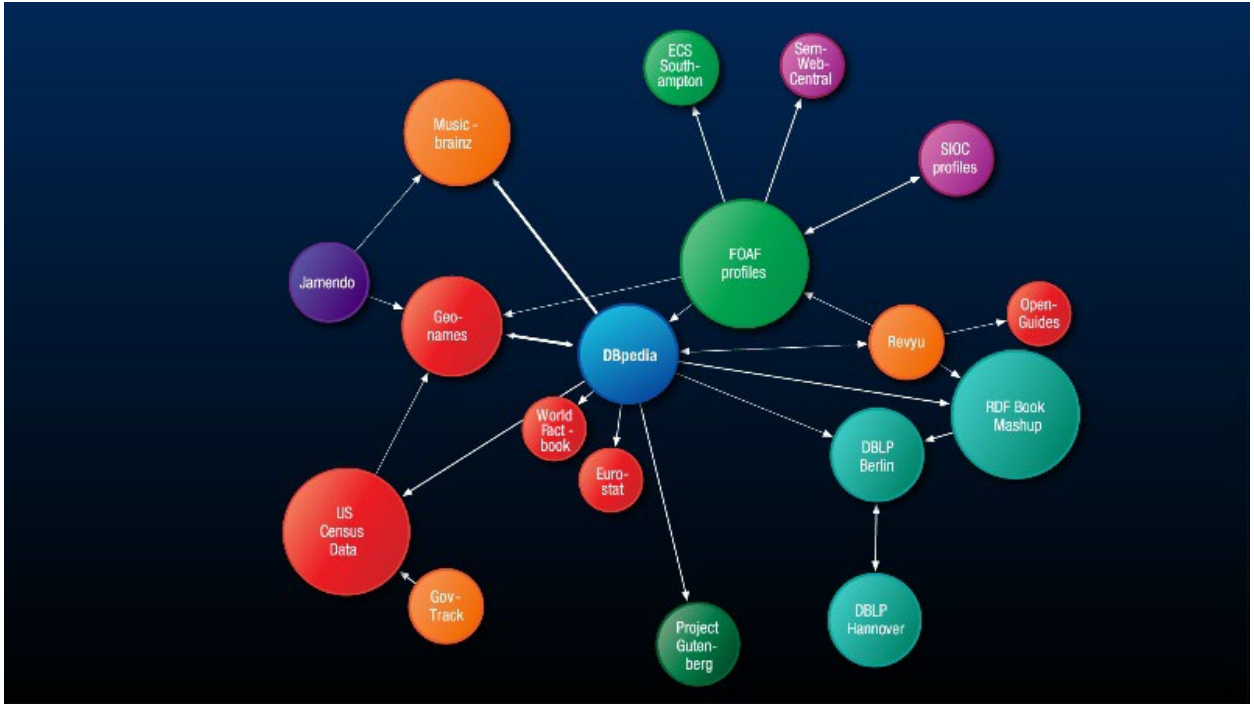
Publish Linked Data in SPARQL endpoints ready for querying

Publishing Linked Open Datasets:
Story Starting 2005



(Tim Berners-Lee)

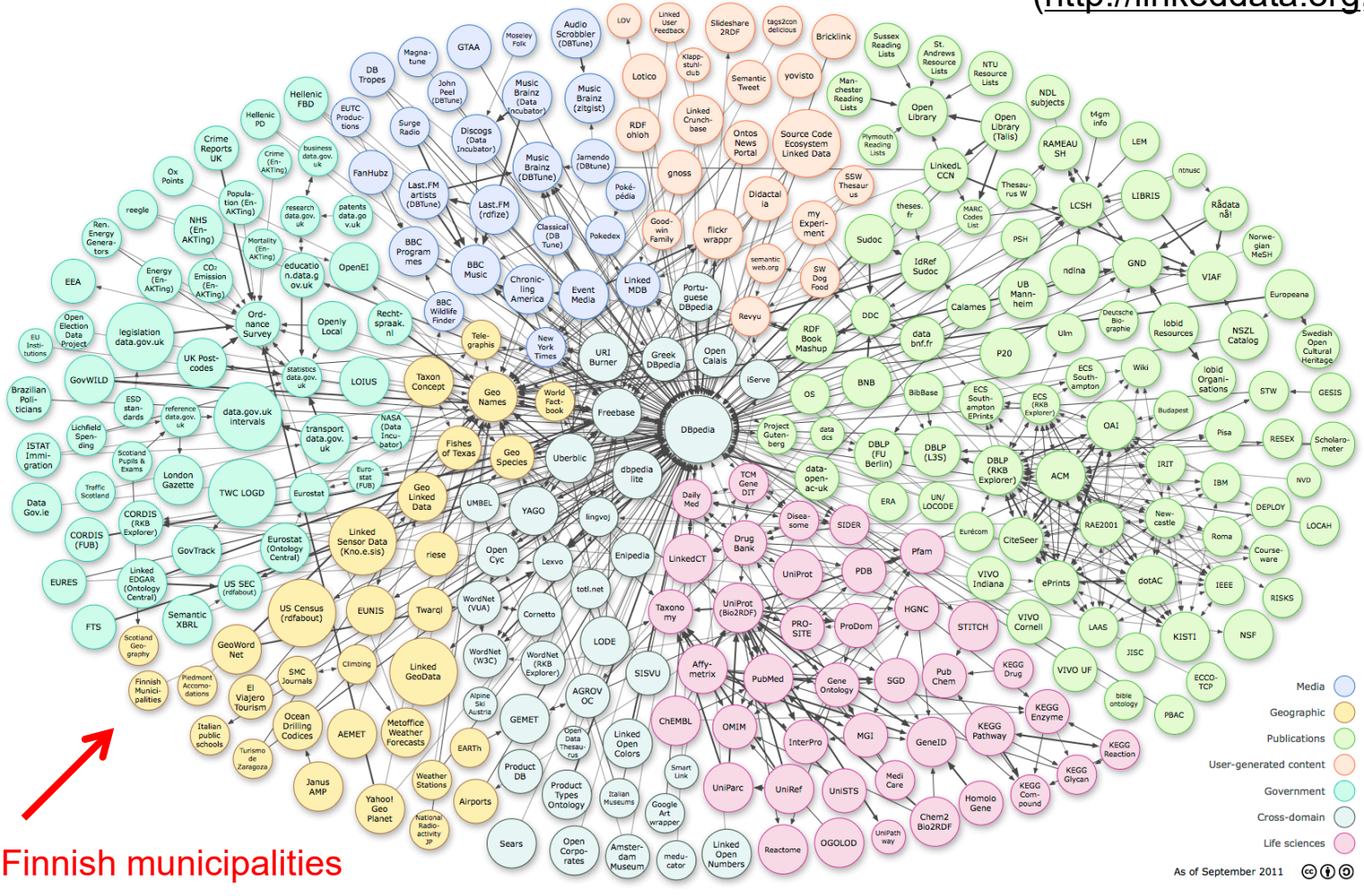
New Datasets Linked with Others



(Tim Berners-Lee)

Linked Open Data cloud 2011

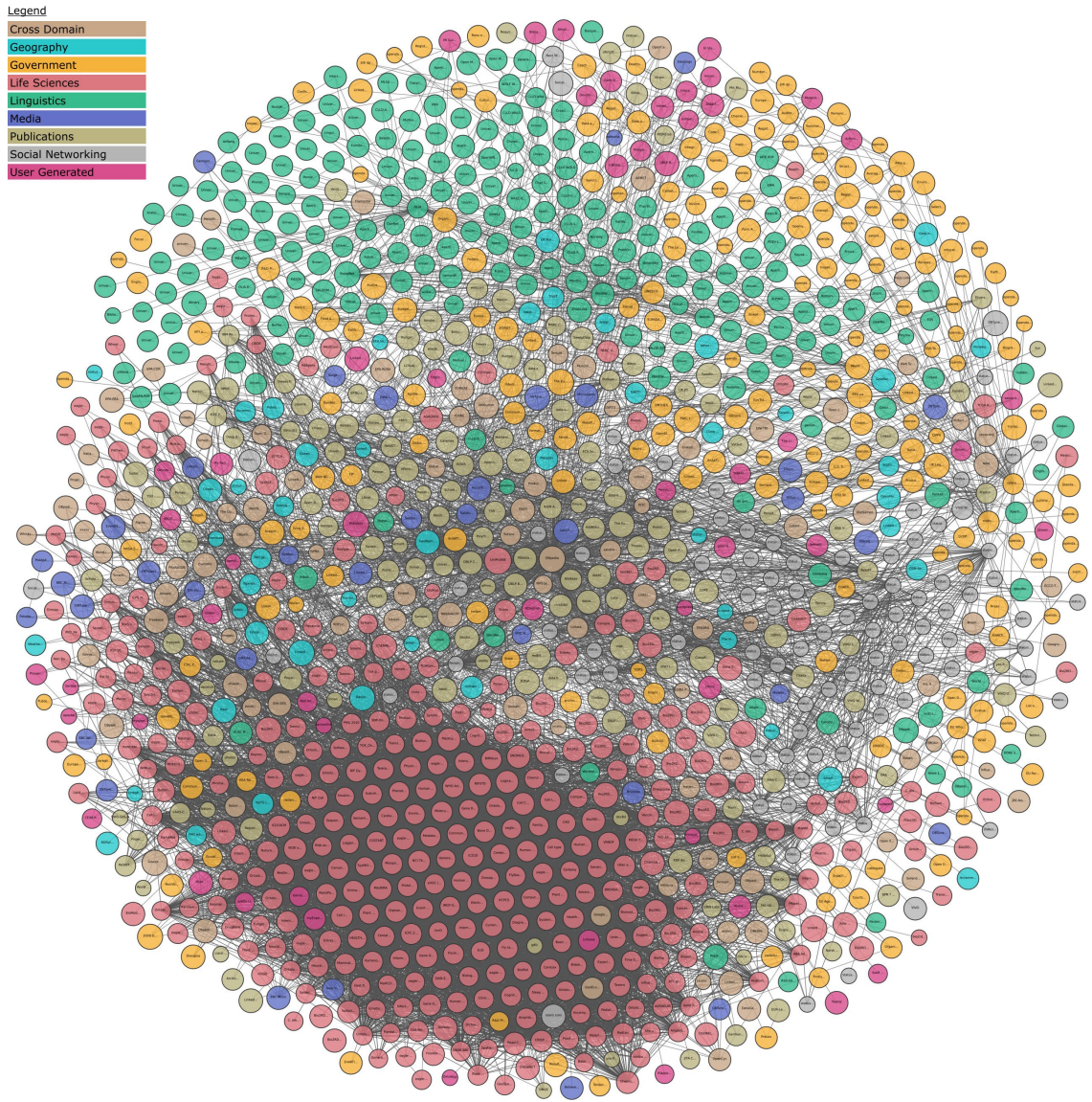
(http://linkeddata.org, 2011)



onki.fi: Finnish municipalities

2021

https://lod-cloud.net/



The Linked Open Data Cloud from lod-cloud.net

There are SPARQL endpoints around

https://www.w3.org/wiki/SparqlEndpoints

Page | Discussion | Read | View source | View history | Search

SparqlEndpoints

Main Page
Browse categories
Recent changes

Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information

see also: SPARQL

In addition to the list below, Mondeca provides a SPARQL endpoint uptime service which monitors the availability of all SPARQL endpoints that are cataloged in CKAN. A similar service is provided by Vienna University.

Currently Alive SPARQL Endpoints

(alphabetical. let's avoid PoorMansHypertext and in-your-face URIs, please)

Project	status	SPARQL endpoint	Webform	comment
Wikidata	(2017-02-23) alive	endpoint	GUI	See also SPARQL federation input
BBC Programmes and Music	(2010-06-29) alive	endpoint	Ajax based Visual Query Builder	Powered by OpenLink Virtuoso; also supports Faceted Browsing and Exploration
Bio2RDF	(2010-01-07) alive	List of 40 SPARQL endpoints	n/a	uses OpenLink Virtuoso
BioGateway	(2010-01-07) timeout	endpoint	webform	BioGateway provides many parameterizable SPARQL queries, both biological as ontological, on RDF graphs that were optimized for querying. The graphs have relational closures. Empowered by OpenLink Virtuoso.
BBC Backstage (HP Labs)	(2010-01-07) server not responding	endpoint	webform	uses joseki 3
BBC John Peel sessions from DBTune (Centre for Digital Music, Queen Mary, University of London)	(2010-01-07) alive	endpoint	n/a	dbtune aims to gather and interlink music-related information.
BBC playcount data from DBTune (Centre for Digital Music, Queen Mary, University of London)	(2010-01-07) alive	endpoint	n/a	dbtune aims to gather and interlink music-related information.
DailyMed	(2010-01-07) alive	endpoint		a **D2R** endpoint
data.gov	(2010-05-22) alive	endpoint	webform	uses OpenLink Virtuoso
data.gov.uk	(2010-02-04) alive	endpoint	webform	The **data.gov.uk** endpoint
DBLP Bibliography Database published through D2R Server (Freie Universität Berlin)	(2010-01-07) alive	endpoint	webform (Maybe only Firefox)	The DBLP database provides bibliographic information on major computer science journals and conference proceedings.
DBpedia (University of Mannheim, Universität Leipzig, OpenLink Software)	(2010-01-07) alive	endpoint	SNORQL webform (Firefox/Safari/Opera); Ajax based Visual Query Builder	dbpedia.org is a community effort to extract structured information from periodic Wikipedia dumps and to make this information available on the Web. It is served to the public via a live instance of OpenLink Virtuoso, and also offers Faceted Browsing and Exploration
DBpedia-live (Universität Leipzig, University of Mannheim, OpenLink Software)	(2010-01-07) alive	endpoint	webform	Based on, now parallel to, and soon to replace the existing dbpedia.org data sets, DBpedia-Live is constantly updated, based on Wikipedia change-feeds. It is served to the public via a live instance of OpenLink Virtuoso, and also offers Faceted Browsing and Exploration
German DBpedia (AG Corporate Semantic Web, Freie Universität Berlin)	(2008-10-15) alive	endpoint	site	de.dbpedia.org is the German language chapter of DBpedia
DBpedia Live German (AG Corporate Semantic Web, Freie Universität Berlin)	(2008-10-15) alive	endpoint	site	de.dbpedia.org is the German language chapter of DBpedia
Spanish DBpedia (Universidad Autónoma de Madrid, Universidad Politécnica de Madrid, OpenLink Software)	(2011-04-04) alive	endpoint	site	es.dbpedia.org is the Spanish chapter of DBpedia
Diseasome	(2010-01-07) alive	endpoint		a **D2R** endpoint
DoapSpace	(2010-01-07) away	endpoint	webform	This is a highly experimental TurboGears with rdflib triplestore (mysql) SPARQL endpoint.
DrugBank	(2010-01-07) alive	endpoint		a **D2R** endpoint
EEA (European Environment Agency) Semantic				

An example of an RDF production pipeline: Case WarSampo

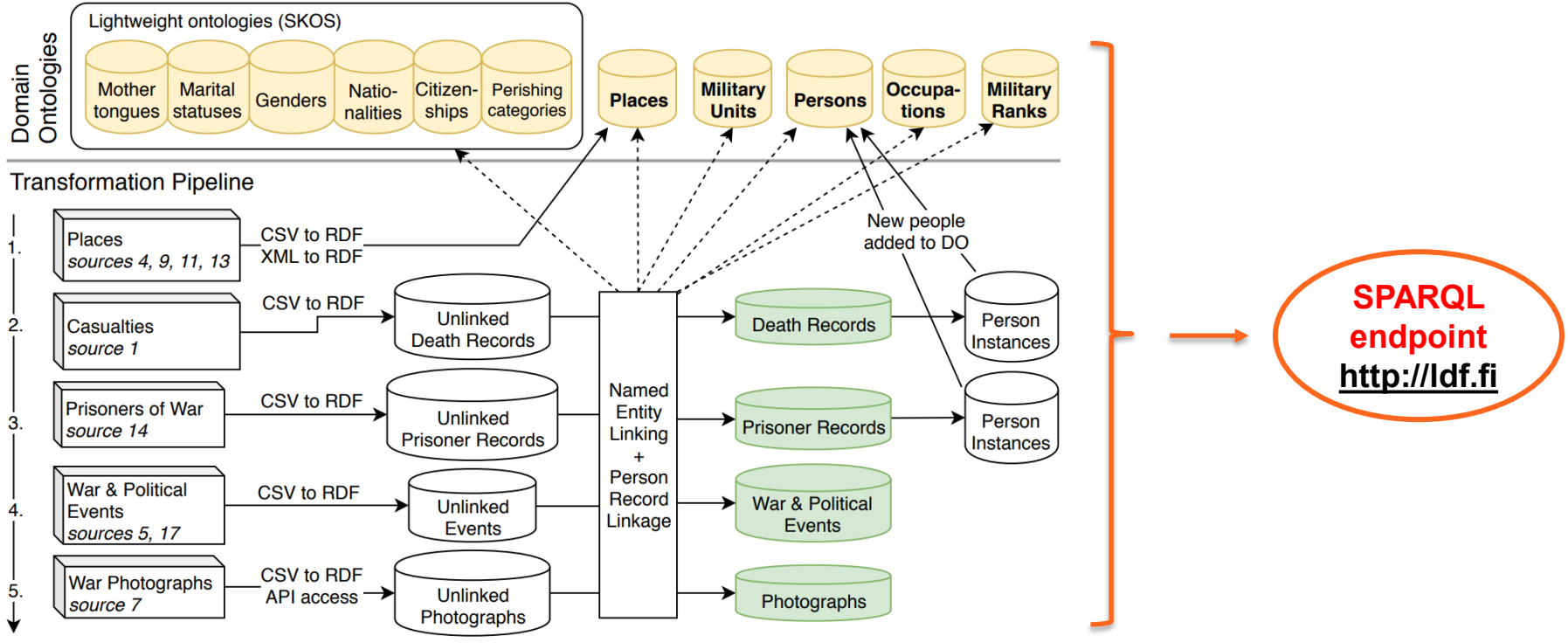
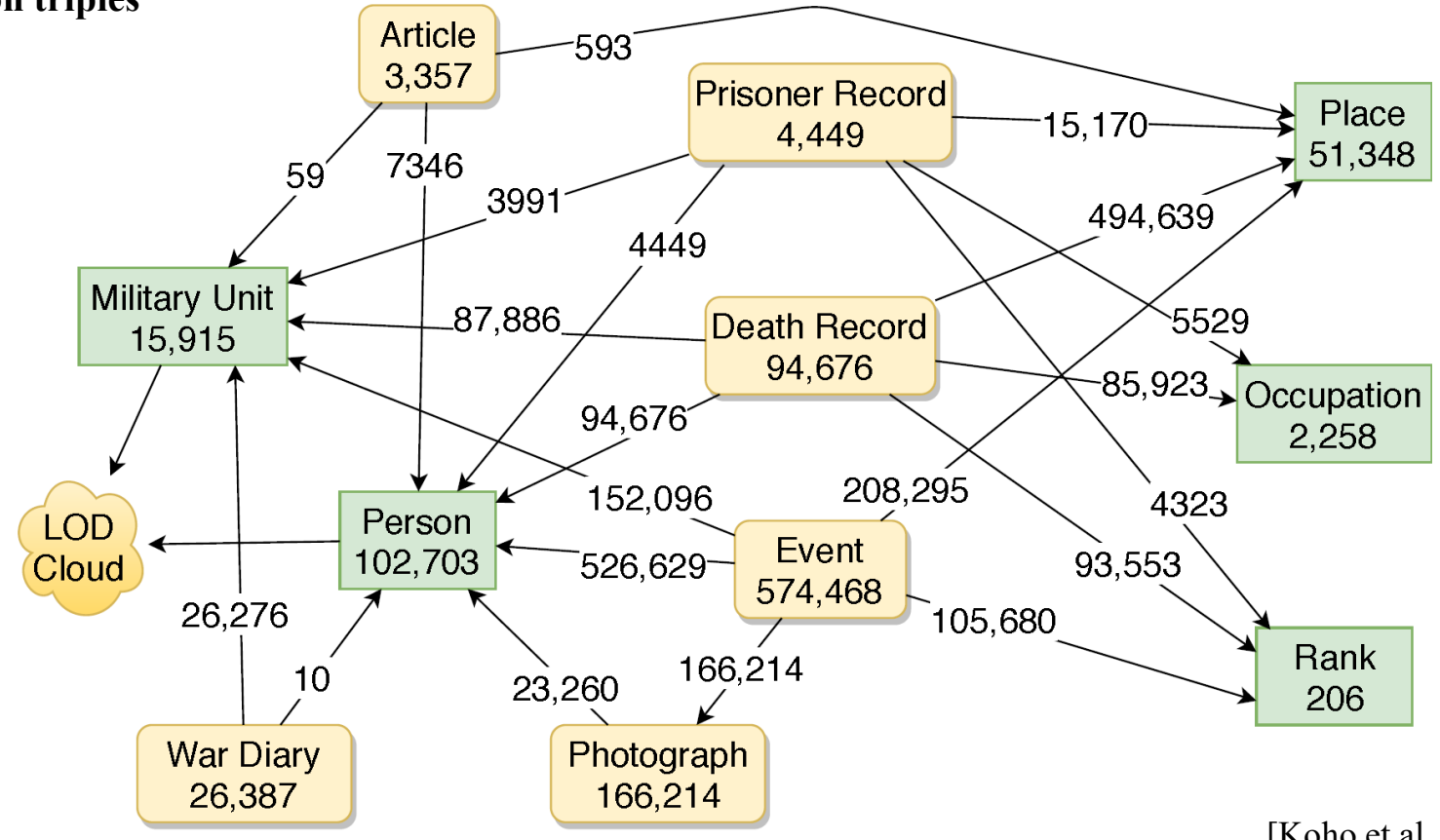


Figure 3. The 5-step WarSampo data transformation process. Dashed arrows represent entity linking, while solid arrows convey data flow.

WarSampo: A Linked Open Data cloud of its own

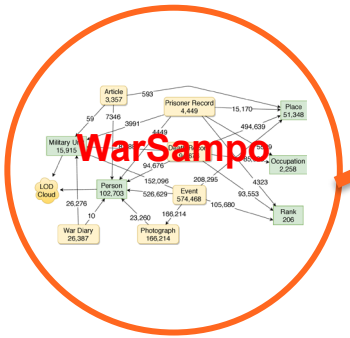
14 million triples

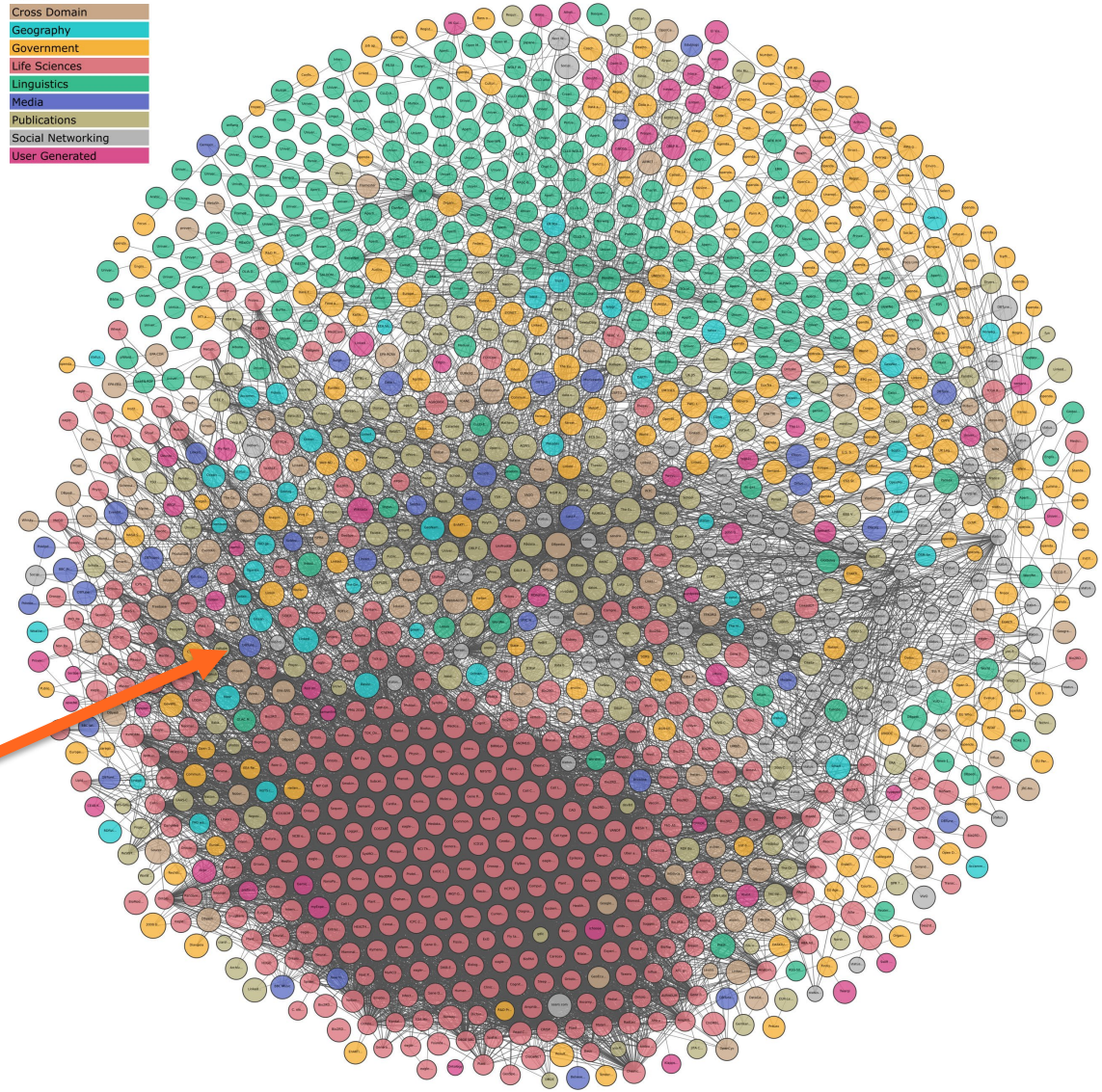


[Koho et al., 2021]

2021

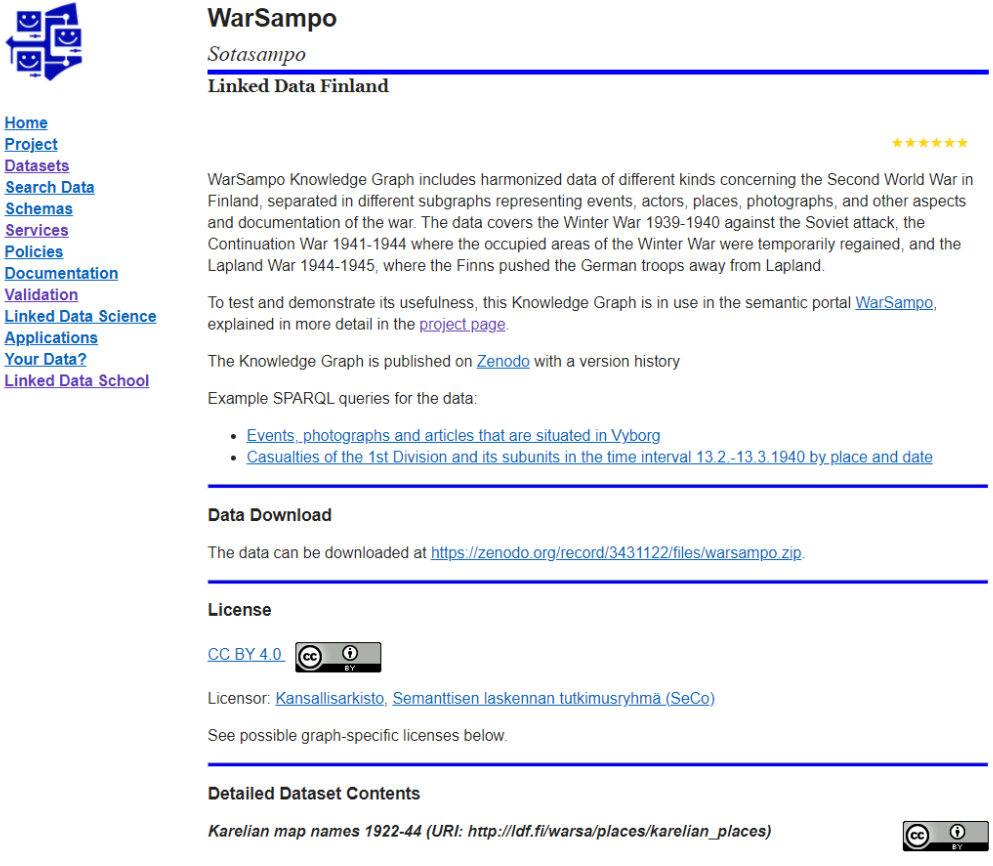
https://lod-cloud.net/





The Linked Open Data Cloud from lod-cloud.net

WarSampo dataset page & SPARQL endpoint: http://www.ldf.fi/dataset/warsa



Home
Project
Datasets
Search Data
Schemas
Services
Policies
Documentation
Validation
Linked Data Science
Applications
Your Data?
Linked Data School

WarSampo

Sotasampo

Linked Data Finland

★★★★★★

WarSampo Knowledge Graph includes harmonized data of different kinds concerning the Second World War in Finland, separated in different subgraphs representing events, actors, places, photographs, and other aspects and documentation of the war. The data covers the Winter War 1939-1940 against the Soviet attack, the Continuation War 1941-1944 where the occupied areas of the Winter War were temporarily regained, and the Lapland War 1944-1945, where the Finns pushed the German troops away from Lapland.

To test and demonstrate its usefulness, this Knowledge Graph is in use in the semantic portal WarSampo, explained in more detail in the project page.

The Knowledge Graph is published on Zenodo with a version history

Example SPARQL queries for the data:

- Events, photographs and articles that are situated in Vyborg
- Casualties of the 1st Division and its subunits in the time interval 13.2.-13.3.1940 by place and date

Data Download

The data can be downloaded at https://zenodo.org/record/3431122/files/warsampo.zip.

License

CC BY 4.0

Licensor: Kansallisarkisto, Semanttisen laskennan tutkimusryhmä (SeCo)

See possible graph-specific licenses below.

Detailed Dataset Contents

Karelian map names 1922-44 (URI: http://ldf.fi/warsa/places/karelian_places)

Information about Linked Data Available and Use Cases

Aalto University
School of Science

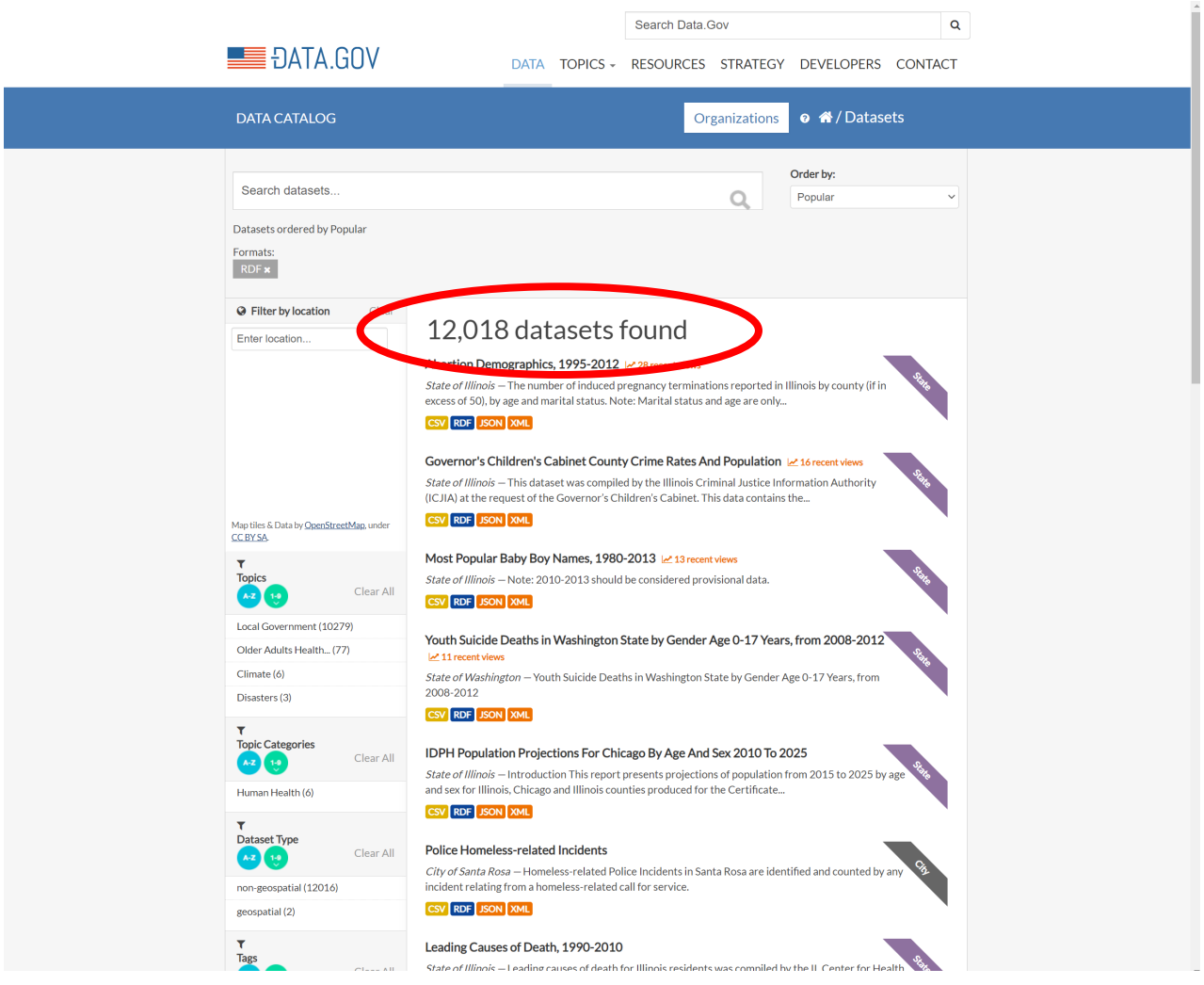
Department of
Computer Science

SeCo

Use cases

W3C's Working Group Note: Data on the Web Best Practices Use Cases & Requirements

In the USA:

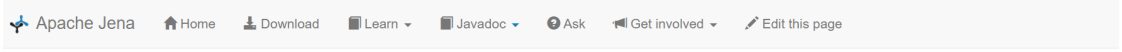


https://catalog.data.gov/dataset?res_format=RDF&_res_format_limit=0

How to create a SPARQL endpoint and applications on top of it by yourself?

Apache Jena Fuseki

https://jena.apache.org/documentation/fuseki2/



DOCUMENTATION / FUSEKI2

Apache Jena Fuseki

Apache Jena Fuseki is a SPARQL server. It can run as a operating system service, as a Java web application (WAR file), and as a standalone server.

Fuseki comes in in two forms, a single system "webapp", combined with a UI for admin and query, and as "main", a server suitable to run as part of a larger deployment, including with Docker or running embedded. Both forms use the same core protocol engine and same configuration file format.

Fuseki provides the SPARQL 1.1 protocols for query and update as well as the SPARQL Graph Store protocol.

Fuseki is tightly integrated with TDB to provide a robust, transactional persistent storage layer, and incorporates Jena text query.

Contents

The Jena users mailing is the place to get help with Fuseki.

Email support lists

Download Fuseki with UI

Releases of Apache Jena Fuseki can be downloaded from one of the mirror sites:

Jena Downloads

and previous releases are available from the archive. We strongly recommend that users use the latest official Apache releases of Jena Fuseki in preference to any older versions.

Fuseki download files

INSTRUCTIONS OF USING FUSEKI:

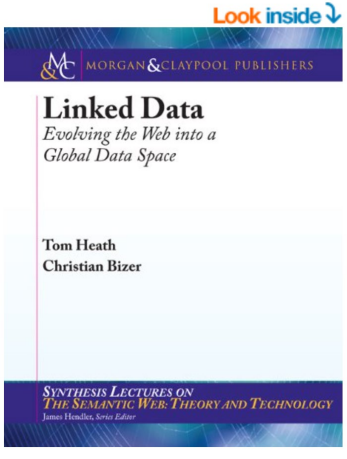
1. Download Fuseki and install it on your machine
2. Start Fuseki
3. You can use it at localhost with your browser
4. Upload an RDF file in the SPARQL endpoint
5. You are ready to query and develop applications!
6. The application can later be published on a web server

More Information – Questions?

Semantic Web & Linked Data Standards
http://www.w3.org/standards/semanticweb/

T. Heath, C. Bizer: Evolving the Web into a Global Data Space
Free online version: http://linkeddatabook.com/editions/1.0/

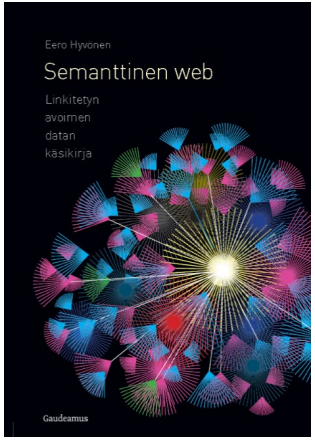
WarSampo project homepage:
https://seco.cs.aalto.fi/projects/sotasampo/en/



In English

2011

https://www.amazon.com/Linked-Data-Evolving-Global-Space-ebook/dp/B009KC1YM2



In Finnish

2018

https://www.gaudeamus.fi/semanttinen-web/