

A!

Aalto University
School of Science
and Technology

DataSuomi

Metadata editor and smart search portal for
Open datasets

School of Science and Technology
Department of Media Technology

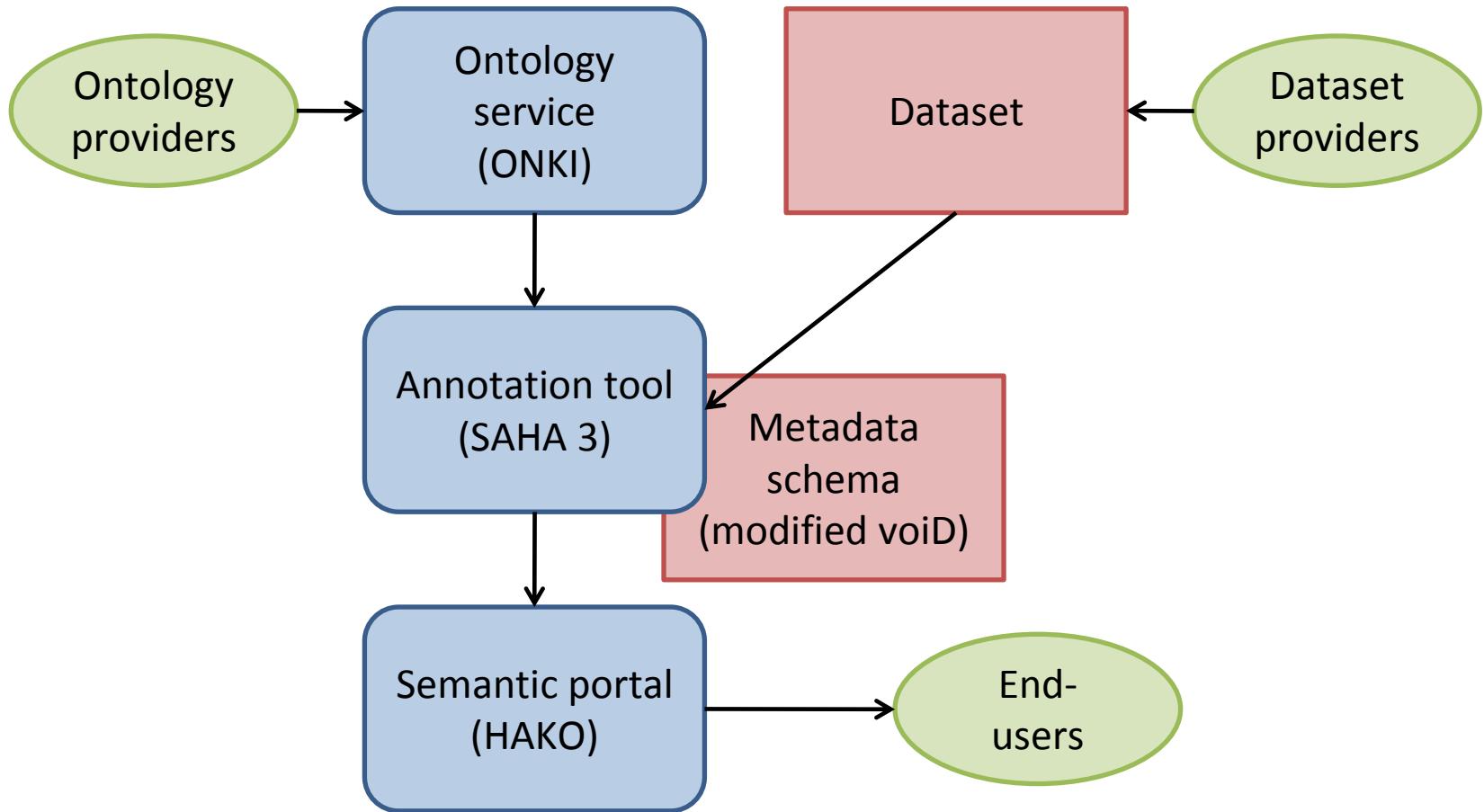
University of Helsinki
Department of Computer Science



Introduction to DataSuomi

- DataSuomi is a tool for editing metadata on and searching for open datasets
- There are two main components to DataSuomi
 - SAHA 3 is used for creating and editing metadata on datasets
 - HAKO portal allows for the publishing of the annotated datasets in a searchable, view-based portal on the web
- The metadata schema used in the annotations is a modified version of the Vocabulary of Interlinked Datasets
 - Allows also non-RDF datasets

Overview of the publication process 1/2

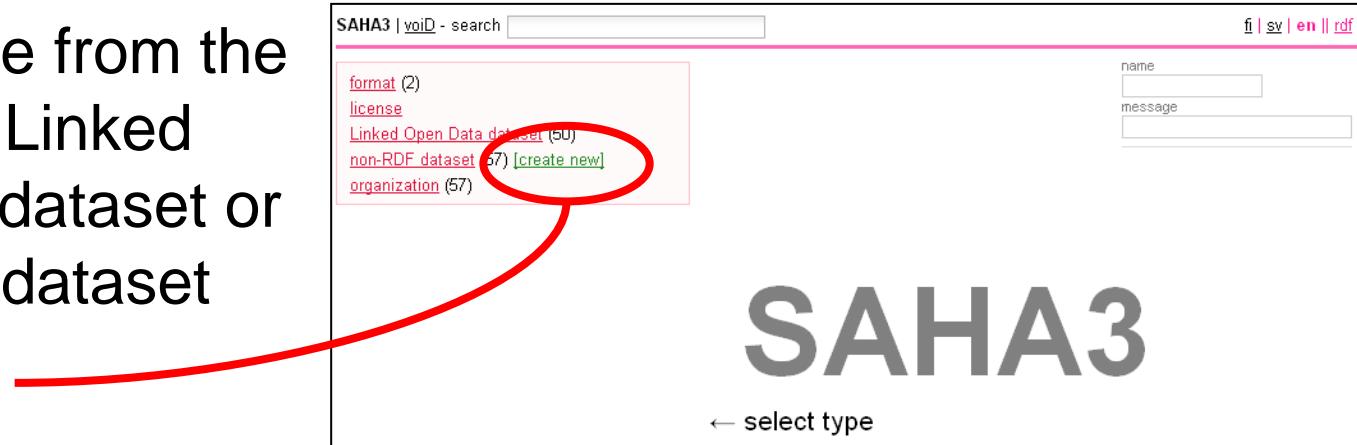


Overview of the publication process 2/2

- First, a dataset is published
- Metadata about the dataset recorded using an annotation tool, SAHA 3 in this case, according to a metadata schema, modified voiD
- Interoperability between annotations is achieved through the use of shared ontologies provided by an ontology service
- Finally, the metadata about the datasets is published in a smart portal, view-based search engine HAKO in this case

Instructions on how to annotate and publish a dataset using DataSuomi

- First, choose from the left either a Linked Open Data dataset or a non-RDF dataset



Instructions on how to annotate and publish a dataset using DataSuomi

- A new dataset annotation has been created
- There are two kinds of values that can be input
 - Literals, where you can write free text
 - References, that refer to RDF resources defined somewhere else
 - Reference fields feature auto-completion (see next slide)

The screenshot shows a web-based form for annotating a dataset. The top navigation bar includes links for 'SAHA3 | void - search' and language options 'fi | sv | en || rdf'. Below the header, the URL 'http://seco.tkk.fi/saha3/u82f36fa4-14d1-4dda-aa7a-79bb024275cf' is displayed. The main content area is titled 'Linked Open Data dataset: u82f36fa4-14d1-4dda-aa7a-79bb024275cf' with a 'remove' link. The form consists of several fields:

title	<input type="text" value="en"/> add new literal
creator The person or organization responsible for the dataset	select reference
homepage	<input type="text"/> add new literal
subject key concepts related to the dataset	select reference
description	<input type="text" value="en"/> add new literal
is published under license	select reference
language of the dataset	select reference
format of the dataset The format in which the dataset is available (e.g. RDF, PDF, Word document etc.)	select reference
has data dump at Announcement of an RDF dump of the dataset.	<input type="text"/> add new literal
has a SPARQL endpoint at	<input type="text"/> add new literal
has an URI look-up endpoint at	<input type="text"/> add new literal

Two specific fields are highlighted with red circles: 'title' and 'subject'. Red arrows point from the 'free text' and 'auto-completion' bullet points in the list above to these respective fields.

Instructions on how to annotate and publish a dataset using DataSuomi

- For example, one of the subjects of BBC Music dataset is music, which can be found by writing ‘mus’ in the subject field and choosing music from the drop-down menu

The screenshot shows the DataSuomi annotation interface for a BBC Music dataset. The 'subject' field is highlighted with a red circle, showing a dropdown menu with suggestions from the koko ontology. The suggestions include 'music', 'music academies', 'musical analysis', 'musical colleges', 'musical exercise', 'musical instruments', 'musical notation', 'musicals', 'musical talent', 'music awards', 'music business', 'music classes', and 'music clubs'. The suggestion 'music' is currently selected.

SAHA3 | voiD - search fi | sv | en | rdf

http://seco.tkk.fi/saha3/u261a6ff1-e519-4e00-b0a0-6f576fba1f4d

name
message

Linked Open Data dataset: BBC Music
[view] | [rdf] | [config] | [remove]

title add new literal
en add
[remove] (i) BBC Music

creator select reference
BBC
[remove] BBC [edit]

homepage add new literal
http://www.bbc.co.uk/music add
[remove] http://www.bbc.co.uk/music

subject key concepts related to the dataset
select reference
<span onMouseOver="showExternalResou
results from koko ontology (showing top 15 of 48 results)
music
music
music academies
musical analysis
musical colleges
musical exercise
musical instruments
musical notation
musicals
musical talent
music awards
music business
music classes
music clubs

description

is published under license

Instructions on how to annotate and publish a dataset using DataSuomi

- For the creator, format and license fields you can click the down-arrow on the field and choose “inline”
- A new annotation frame opens where you can describe the reference directly

SAHA3 | voiD - search fi | sv | en | rdf

http://seco.tkk.fi/saha3/u261a6ff1-e519-4e00-b0a0-6f576fba1f4d

name
message

Linked Open Data dataset: BBC Music
[view] | [rdf] | [config] | [remove]

title add new literal
en add
[remove] (fi) BBC Music

creator select reference
The person or organization responsible for the dataset
[close]
title add new literal
homepage add new literal
description add new literal
[remove] BBC [edit]
[remove] u98266955-0-2c147-2c-a515-91e34b7011ae [edit]

homepage add new literal
[remove] http://www.bbc.co.uk/music add

subject key concepts related to the dataset
select reference
[remove] art music [edit]
[remove] blogs [edit]
[remove] entertainers [edit]
[remove] jazz [edit]
[remove] music [edit]
[remove] news [edit]
[remove] popular music [edit]
[remove] rock'n'roll [edit]
[remove] äänilevyt [edit]

description add new literal
en add

Instructions on how to annotate and publish a dataset using DataSuomi

- HAKO enables the searching of datasets
- Free text search field is in the upper left corner
- You can also choose facets from the left to narrow down your search

Hako - Faceted Search Engine [reset]

VoID

Search

Format of the dataset

RDF

[remove] Economy

[remove] English language

[remove] RDF

Results 3

Subject

Bureaus

Cabinets

Culture

Databases

Decision-making

Economy

Films

Geography

Human population

International trade

Literature

Music

Natural resources

Persons

Politics

Public administration

Science

Society

Sport

State institutions

States

States of a federation

Statistics

Television

Language of the dataset

English language

CIA Factbook (Linked Open Data dataset)

data.gov (Linked Open Data dataset)

Freebase (Linked Open Data dataset)

Links

- DataSuomi
 - SAHA 3: <http://demo.seco.tkk.fi/saha3sandbox/voID/index.shtml>
 - HAKO: <http://demo.seco.tkk.fi/saha3sandbox/voID/hako.shtml>
- Research article:
<http://www.seco.tkk.fi/publications/submitted/frosterus-hyvonen-saha-void-2010.pdf>