



Challenges in managing semantic annotations in harvested research objects in a national CRIS context

Tommi Suominen - <https://orcid.org/0000-0002-4382-7425> , CSC - IT Center for Science
Joonas Kesäniemi - <https://orcid.org/0000-0002-3770-0006> , CSC - IT Center for Science
Katja Mankinen - <https://orcid.org/0000-0001-7357-9648> , CSC - IT Center for Science



Contents

- Research.fi
- Semantic annotation
- Word clustering
- Challenges in maintaining annotations, when the background changes

The Finnish Research Information Hub

- The National Research Information Hub project was launched in 2017
- Owned and financed by the Ministry Culture of Finland that has authorised the implementation to CSC – IT Center for Science Ltd.
- Information is publicly and freely available in www.research.fi launched in 2020



Research.fi

Check
it out!



Aims of the Finnish Research Information Hub

R&D analytics
Visualizations
Business
Intelligence
Landscape

To describe the Finnish
research ecosystem and
act as an information
base for science policy

To help to find Finnish
research and researchers

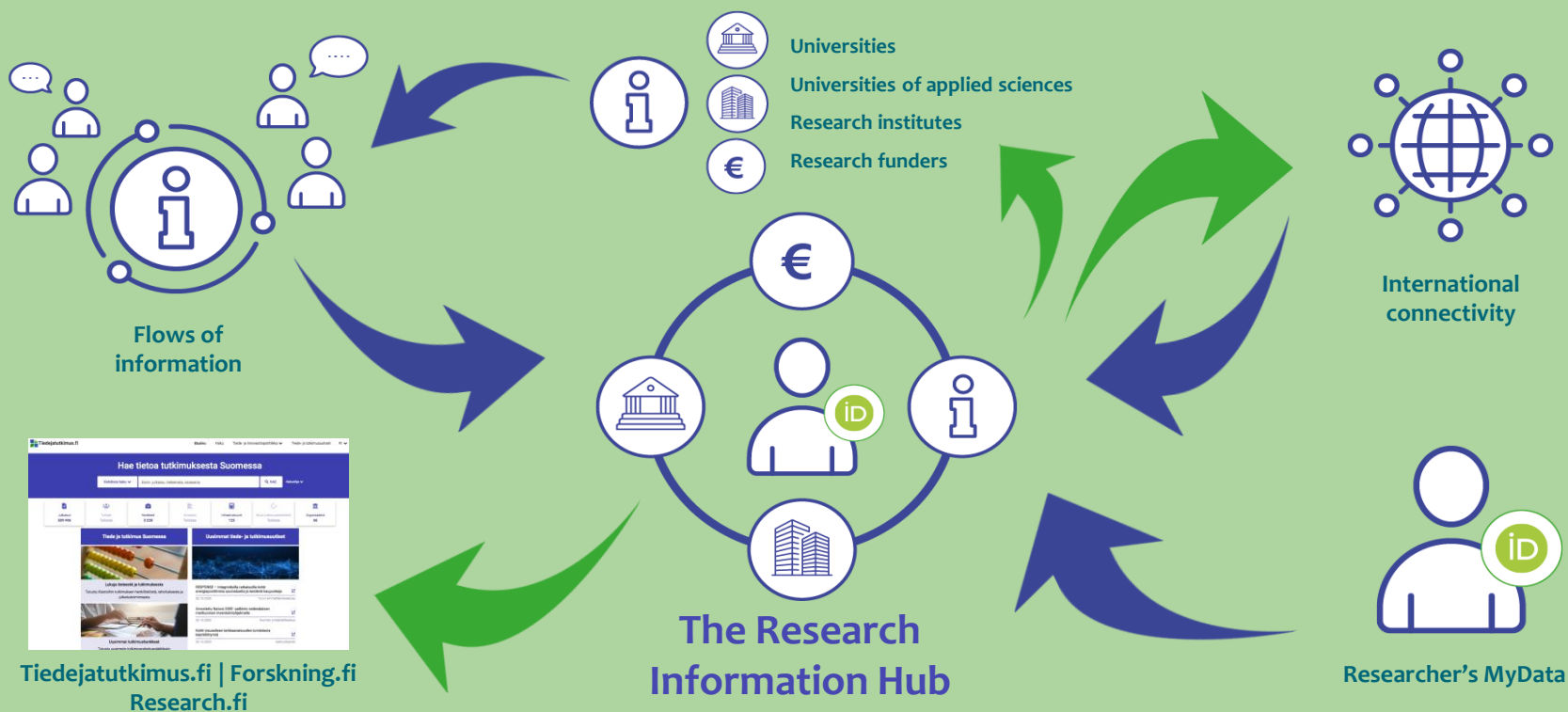
Research.fi portal
Semantic search
Linked data

APIs
Shared data
model

To support the reuse of
existing research
information in other
services and systems

To support the merit of
researchers and increase
the visibility of their
research output and
activities

Researcher's Profiles
Traditional research outputs
Other research activities



Why

1. The details of researchers, publications, research datasets, research projects and research infrastructures will soon be available in one place.
2. There will be less reporting and administrative work as the data is available in a single location and information flows freely between services.
3. Results of both publicly and privately funded research will be openly accessible.
4. This will benefit researchers, research organizations, funding agencies, public administration and citizens.

Search for information on research in Finland

Search target ▾

For example, publication, field of science, keyword

Q SEARCH

Search help ▾

Publications
629 359People
Coming soonProjects
8 695Research data
11 198Infrastructures
156Other research activities
Coming soonOrganizations
77

Science and research in Finland



Figures on science and research

Read statistics on research personnel, funding and publishing.



Latest research projects

Latest science and research news

Savonia-artikkeli: Podcastit osana työikäisten mielenhyvinvoinnin edistämistä



11.05.2022

Savonia-ammattikorkeakoulu

Savonia-artikkeli: Arvokkaan kuoleman kohtaaminen vaatii osaamista



11.05.2022

Savonia-ammattikorkeakoulu

Savonia-artikkeli: iPana Äitiys -järjestelmän käytettävyyttä tukee kättilön työtä, mutta vain osittain



11.05.2022

Open funding calls

Stipendiohjelmat maisteri- ja tohtoritason opintoihin/tutkimukseen Yhdysvaltoihin lukuvuodelle 2023-24!

Hakuaika päättyy **22.05.2022**
Fulbright Finland Foundation

Tuki jatko-opintoihin ulkomailla

Hakuaika päättyy **31.05.2022**
Jorma ja Märtha Sihvolan säätiö sr

Liike-elämän ja koulutuksen välinen yhteistyö

Hakuaika päättyy **31.05.2022**
Suomen Liikemiesten Kauppaopiston Säätiö sr

GIVE FEEDBACK!

Semantic concept annotation and search

Implementation by: Joonas Kesäniemi



Relationships between linkable items (research objects)

- Subject – predicate –object (triplet)
- Depending on the subject and object in question, different predicates will apply. This is fine in the linked data world. In the graph sense, this implies more than two linkable items being simply connected.

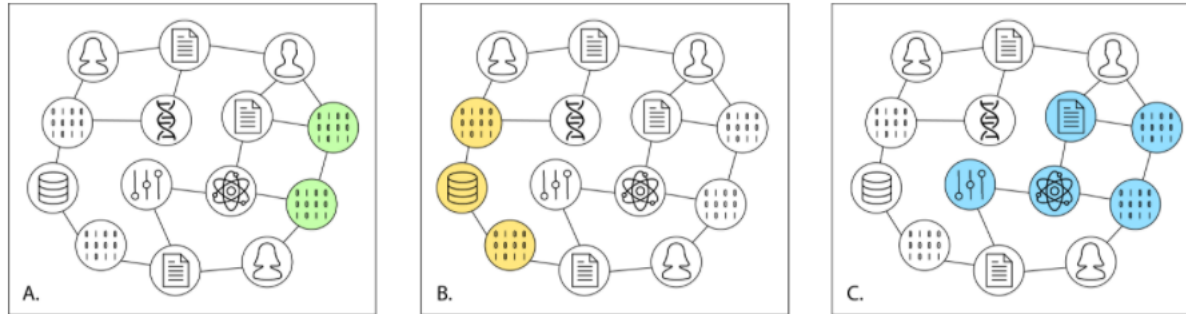


Fig 1. A schematic representation of the PID graph with digital objects connected by PIDs, showing three use cases: A: Different versions of software code, B: Datasets hosted by a particular repository, C: All digital objects connected to a research object.

Fenner, M., Aryani, A., 2019. Introducing the PID Graph
 — FREYA [WWW Document].
<https://doi.org/https://doi.org/10.5438/jwvf-8a66>

Also links from research objects to concepts form triplets

- Can also be used in reverse, research objects related to a specific concept:

<https://sanastot.suomi.fi/concepts/3bdbcac2-e57f-49c1-b104-e37eff042834/concept/8395b72f-7015-4d77-b0b5-e1699d79c867>

- Semantic research object

search

- Navigating research outputs using concept relationships

Research management vocabulary

Valid Information domain: General information and administrative services Education and training Organization: Finnish Association of Research Managers and Advisors

Content language ▼

CONCEPTS TERMINOLOGY

ALPHABETIC HIERARCHICAL COLLECTION

tutkimus 🔍

- anonymisation of research data
- applied research
- basic research
- competed research funding
- contribution in research
- data-intensive research infrastructure
- digital research data
- equipment-intensive research infrastructure
- granted research funding
- leadership in research work
- national roadmap for research infrastructures
- Principal investigator
- project plan
- pseudonymisation of research data
- research
- research and development (R&D)
- research area
- research community
- research consortium
- research data
- research ecosystem
- research environment

research output

Valid

Preferred term

- FI tutkimustuotos
- EN research output

Definition

- FI yksilöitävissä oleva ilmentymä, joka valitsee tietoa tehdystä tutkimuksesta ja siinä syntyneistä tuloksista

Note

- EN CASRAI term: Outputs

Information representing all the tangible results of the research or scholarly activity of a person or organization.

Concept class

Julkaiseminen ja hyödyntäminen

Concept status

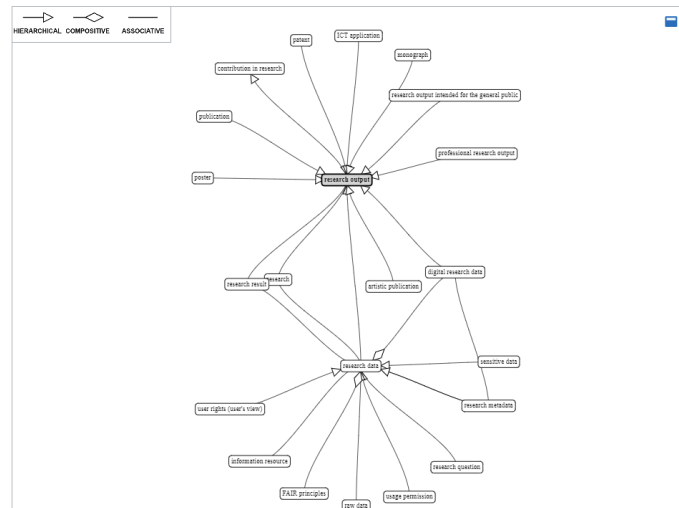
Valid

Broader concept

- contribution in research

Narrower concept

- publication
- artistic publication
- ICT application
- patent
- research data
- digital research data



Semantic annotation in research.fi

- Semantic annotation is based on utilising a known ***ontology*** and an ***annotator*** that has been trained to annotate textual content with concepts.
- These concepts are consequently linked as metadata related to a research object.
- In our context, automated concept annotation uses the YSO-ontology using the Annif-annotator.
- At first, we focus on annotating the funding decisions, based on their titles and descriptions

Semantic search based on the annotations

- A search with the concept "forest management" will find a funding decision, with the term "metsänhoito" (Finnish) used in the description.
- This uses the different language labels of a concept that is a direct hit for the searched concept.
- In the next iteration, the search with "forest management" will find the document with the annotated concept "tree planting", as the two concepts are semantically related.
- Complements the word clustering approach/topic modelling to find funding decisions e.g. by overcoming content language differences.
- First iteration of this implementation is in testing phase
- Next up for development: concept annotation for other research objects

Topic modelling - the word clustering approach

Katja Mankinen, CSC



Topic modelling of research funding decision

- What topics or themes are being researched in Finland or my own organization
- How many projects were funded on climate change/cancer between 2010-2019?
- Which topics/themes received the most funding?
- Unsupervised machine learning methods are used to group text automatically into (often interdisciplinary) topics
- Topic modelling may improve the discoverability of funded research activities (e.g. projects) and offer new insights into the Finnish research field.
- In the Finnish Research Information hub, approx. 30% of the funding decisions are missing information on the field of science that they concern.
- The first iteration of this functionality is in production

Funding decision topic modelling



Data

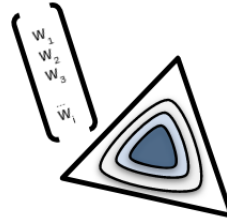
Approx . 7000 decisions (title, description, keywords and field of science)



Text preparation

Certain general concepts, frequently occurring non semantic words are removed, language of the description is detected

The text is converted into numerical form (word and document vectors) to facilitate processing



Topic modelling

Using and comparing different methods (contextualized topic models, top2vec, BERTopic, hierarchical stochastic block model) to filter potential topics, based on the relationships between words in the descriptions

Topic = a cluster of words, which have a semantic relationships of some kind

Based on empirical experimentation, 92 topics were selected , and each funding decision is classified into the most significant topic



Small, large, old , rising, interdisciplinary ...
→ topics may evolve and new topics may appear
→ the model is updated, when new data is available

Filtering based on identified topic

Filters (101)

Close

Topic

- 3d printing, printing X SMEs, management, growth, entrepreneurship X aerosol, atmospheric X
- agriculture, farming, dairy X air quality X animals X arctic region X art, culture X
- artificial intelligence, machine learning X bacteria, antibiotic infections X
- batteries, lithium batteries, battery X biodiversity, species, ecosystems, evolution, ecology X
- biomass, bioeconomy, biorefinery, biofuels, biogas X
- biomass, cancer, bioeconomy, biorefinery, biofuels, biogas X brain diseases, Alzheimer, dementia X
- business, investments X cancer X cancer, arctic region X cancer, art, culture X
- cancer, business, investments X cancer, mining, minerals X carbon, methane, carbon footprint X
- cardiology, cardiac, ventricular, myocarditis X catalysts X chemicals, pharmaceutical, risks X
- circular economy X climate change, adaption X cloud computing, edge computing X
- computer vision, graphics, face recognition X construction engineering, wood construction X
- consumer, customer X covid, corona virus, corona, pandemic X
- crisis preparedness, disaster, emergency X cybersecurity, encryption, cryptography X
- diabetes, metabolic diseases X disorders, mental health, psychiatry, depression, psychotherapy X
- drones, unmanned aerial vehicles X drug, drug delivery, drug development X education X
- energy, electricity, power X eyes, eye diseases, glaucoma X food, nutrition X
- forestry, wood, timber X fuels, combustion X games, gaming X
- gender, feminism, equality, discrimination X health care X heating X history, archaeology X
- imaging, biomedicine X internet of things, iot, devices, wearable electronics X

15

Projects - 8 695

Show as Image

Filter results

Starting year 1

+ Organization 1

Funder 1

+ Funding instrument 1

+ Field of science 1

- Topic 1

Keywords 1

Themes 1

Academy of Finland's research fields 1

Identified topic 1

Search

☐ Select all

☐ other 795

☐ health care 623

☐ manufacturing, production 297

☐ education 291

☐ sustainable ecosystems, sustainable 272

Displaying results 11 - 20 / 8695

10 results / page

What project information is included in the service?

PROJECT NAME	FUNDER	RECIPIENT	STARTING YEAR
 Kun ihminen lakkasi olemasta eläin. Eläin, eläimellisyys ja ihminen-eläin-ero sosiologiassa ja antropologiassa ennen ja jälkeen toisen maailmansodan	Kone Foundation	Salla Tuomivaara, University of Turku	2022
 Making Meaning: A study in foundational semantics	Kone Foundation	Jaakko Reinikainen, Tampere University	2022
 Non-Abelian anyons for topological qubits: TOBITS	Academy of Finland	Manohar Kumar, Aalto University	2022
 Governing Transition to a Circular City: Engaging public and private stakeholders in the circular economy	Academy of Finland	Pekka Valkama, University of Vaasa	2022
 Gallium Oxide Fabrication with Ion Beams	Academy of Finland	Flyura Djurabekova, University of Helsinki	2022
 REcycling of spent Li-ion batteries and end-life photovoltaic panels: From the development of metal recovery processes to the implementation of a START-up (RESTART)	Academy of Finland	Benjamin Wilson, Aalto University	2022
 Lasten ja nuorten oikeus hyvään tulevaisuuteen heidän itse tutkimaan: filosofoiva kasvuyhteisö ja sukupolvien välinen oikeudenmukaisuus	Kone Foundation	Tuukka Tomperi, Tampere University	2022
 Kansakunta - sarja siitä, miten Suomi hajosi ja koottiin uudelleen	Kone Foundation	Kaarina Hazard	2022
Assessment and improvement possibilities for social sustainability of electric vehicle battery material production	Maj and Tor Nessling Foundation	Anni Orola	2022
Finnish Quantum-Computing Infrastructure	Academy of Finland	Mikael Johansson, CSC – IT Center for Science	2022

Visualization

Distribution of projects by year ⓘ Additional information

Funding instrument ⓘ

Field of science ⓘ

Topic ⓘ

Keywords ⓘ

Themes ⓘ

Academy of Finland's research fields ⓘ

Identified topic ⓘ

Search

☐ Select all

☐ other

☒ health care

☒ manufacturing, production

☒ education

795

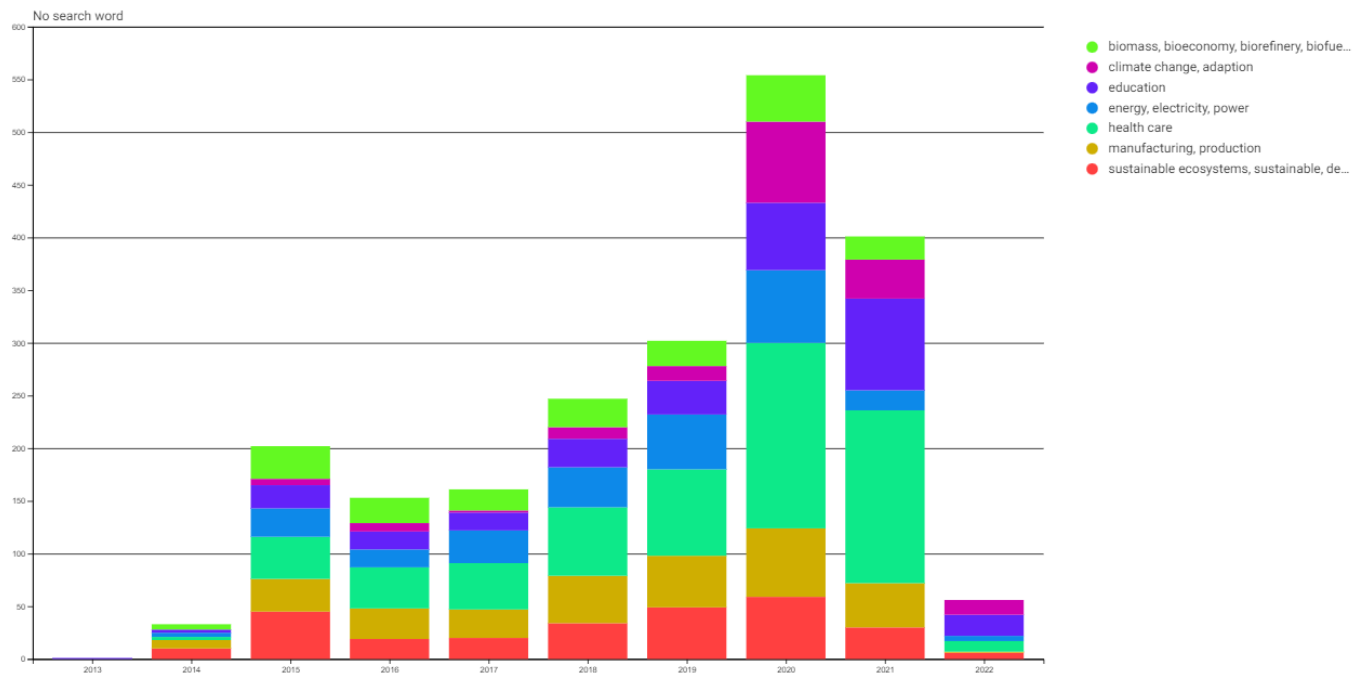
623

297

701

Clear filters

No search word



Year	health care	education	climate change, adaption	energy, electricity, power	biomass, bioeconomy, biorefinery, biofuel	manufacturing, production	sustainable ecosystems, sustainable development
2013	0	0	0	0	0	0	0
2014	10	10	10	10	10	10	10
2015	40	30	20	20	20	20	20
2016	30	20	10	10	10	10	10
2017	40	30	20	20	20	20	20
2018	60	40	20	20	20	20	20
2019	80	50	20	20	20	20	20
2020	120	70	40	40	40	40	40
2021	100	60	30	20	20	20	20
2022	10	10	10	10	10	10	10

- biomass, bioeconomy, biorefinery, biofuel
- climate change, adaption
- education
- energy, electricity, power
- health care
- manufacturing, production
- sustainable ecosystems, sustainable development

Number of projects	Percentage of the number of projects
Amount granted	Percentage of the amount granted

Limitations on topic modelling

- All the funding decisions have not been classified to a topic
- All decision are not necessarily tightly connected to the classified topic
- Significant topics may remain undetected/clustered
- Limitations both in the data and in the methods
 - Small and multilingual dataset
 - Very specific topics of research
 - Some decisions have just a title and/or a very short description
 - The topic modelling is based on unsupervised machine learning, small parameter changes may produce different results
- Topic modelling is suitable for mapping research themes on a large scale, but not for determining exact sizes of individual topics.

Challenges in maintaining annotations, when the background changes



Challenges for topic modelling

- As the body of funding decisions defines the clusters, new funding decisions that come in may change the clusters
 - By redefining existing cluster boundaries
 - Introducing a new cluster (and consequently closing another one, as the number of clusters is fixed)
- How to deal with the temporal evolution of clusters? Useful to track historical cluster assignment? For discussion on the “Dynamics of topics” see e.g. [Mankinen & Leino 2021](#)

Challenges in annotation management

- Semantic annotation is based on utilizing a known **ontology** and an **annotator** that assigns concepts as metadata to a research object.
- The annotator is trained to annotate specific ontology concepts; but if the ontology changes e.g. through the introduction of new concepts, the annotator is not trained for these new concepts.
- Changes in an ontology must be incorporated into the annotator, to operationalize the newly added concepts.
- New data as such does not impact the ontology or the annotator.
- However, it may be that new content does not fall into the scope of the ontology, and therefore adequate annotation of content with concepts might not be achieved

The separation of facts and inferred information

- If we have utilized reasoning/inference to create new machine-generated metadata, then if the facts based on which the reasoning was performed, change, the previous conclusions may become invalid. The process of inference needs to be reconducted and the conclusions redrawn.
- This then requires the separation of facts and inferred information, and potentially the linkage of the facts to their conclusions in a traceable manner. (the provenance of the conclusions)
- This would enable re-inference, when we notice one of the dependent facts changed.
- If we are not able to do this, then the recourse is to re-reason the entire body of inferred conclusions



Thank you for your attention!

I am Tommi Suominen. I work at CSC as an information architect.

www.linkedin.com/in/tommi-suominen-935b5a8a

I will be coordinating the FAIRCORE4EOSC Horizon Europe project starting in June 2022. I hope to be talking about this at the next euroCRIS meeting.



facebook.com/CSCfi



twitter.com/CSCfi



youtube.com/CSCfi



linkedin.com/company/csc---it-center-for-science



github.com/CSCfi