

# **Yle meets Annif**

**- an open source tool for automated subject indexing**

Osma Suominen and Pia Virtanen



EBU MDN workshop  
10 June 2020

# Subject indexing

a.k.a. topic indexing, topic assignment

~ tagging

~ multi-label classification

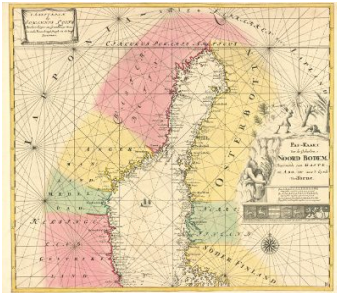
Landsdagen omfattar regeringens proklamation om Finlands fullständiga oavhängig-  
het och ansluter sig till huvudprinciperna i regeringens program för trygghet  
av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka till-  
fölle ett av socialdemokraterna formulerat förslag.

Illegale Handel (Waren und Personen) ist  
et. 2. Weltkrieg ist in der Geschichte  
zusammengefasst. Von der Zeit nach dem  
Zweiten Weltkrieg bis heute ist es  
meistens so, dass die Waren und Personen  
nicht mehr in der Welt sind.

[illegible][illegible][illegible]

Der Herr Abgeordnete hat sich für die  
Arbeit der hiesigen No-Strömchen-  
Kasse nicht nur insofern in Tätigkeit  
gezeigt, als er dieselbe in Tätigkeit  
setzte, sondern er schenkte sich auch an  
der Beschaffung von Arbeitsmaterialien  
für diese Kasse und unterstützte  
dieselbe.

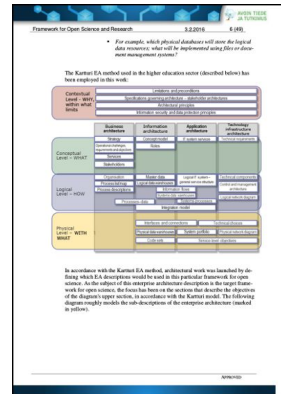
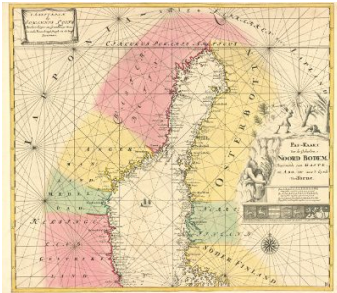
Pria 25 poyal



The Kartari EA method used in the higher education sector (described below) has been employed in this work:

Contextual Level = WHY, where what limits	<ul style="list-style-type: none"> <li>• policies and priorities</li> <li>• legislation, government policies, administrative procedures</li> <li>• financial resources, human resources</li> <li>• standards, security and safety, other programs</li> </ul>			
	Business architecture	Information architecture	Applications architecture	Technology architecture
Conceptual Level = WHAT	Strategy	Information strategy	Applications strategy	Technology strategy
	Business requirements and capabilities	Information requirements and capabilities	Applications requirements and capabilities	Technology requirements and capabilities
Logical Level = HOW	Business processes	Information processes	Applications processes	Technology processes
	Business data	Information data	Applications data	Technology data
Physical Level = WITH WHAT	Business architecture	Information architecture	Applications architecture	Technology architecture
	Business processes	Information processes	Applications processes	Technology processes

In accordance with the Karmali EA method, architectural work was launched by defining which EA descriptions would be used in this particular framework for open science. As the subject of this enterprise architecture description is the target framework for open science, the focus has been on the sections that describe the objectives of the diagram's upper section, in accordance with the Karmali model. The following diagram roughly models the sub-descriptions of the enterprise architecture (marked in yellow).



## YSO - General Finnish ontology

Content language English ✕ Search

A-Z Hierarchy Groups New

- events and action
- objects
  - abstract objects
  - physical objects
  - inanimate objects
    - abacuses
    - adhesive tapes
    - admission tickets
    - amigurumi
    - animal bodies
    - armours
    - articles (inanimate objects)
    - artificial nails
    - artificial organs
    - audience areas
    - balloons
    - bandoles
    - baskets
    - bathtubs
    - bird tables
    - birdhouses
    - bits (bridles)
    - blinds
    - bookmarks
    - booms
    - braids and weaves
    - bridge floors
    - bridles
    - briquets
    - brooms and brushes
    - buckles
    - busses (computing)
    - buttons (clothing)
    - candles
    - candlesticks
    - cardboard
    - cards
    - cart structures
    - celestial bodies
    - ceramics
    - chains (objects)

objects > physical objects > inanimate objects > armours

PREFERRED TERM

**armours** 

TYPE

General concept

BROADER CONCEPT

[inanimate objects](#)

RELATED CONCEPTS

[military uniforms](#)

BELONGS TO GROUP

67 Warfare. Military Technology. Defence. Weapons

IN OTHER LANGUAGES


haarniskat

Finnish

harnesk

Swedish

URI

<http://www.yso.fi/onto/ysop14728> 

Download this concept:

[RDF/XML](#) [TURTLE](#) [JSON-LD](#)

Last modified 5/10/17

EXACTLY MATCHING

[armours](#)

KOKO Ontology

CONCEPTS

[haarniskat \(fi\)](#)

YSA - General Finnish thesaurus

[harnesk \(sv\)](#)

Allärs - General thesaurus in Swedish

Images indexed with the term in Finna 245

[Image](#)



Liivinmaan  
kenraalikuvernööri,



valtaneuvos Pontus De la  
Gardia esittävä



Kustaa Eriksson Vaasa  
taistelussa tanskalaisia



Ruotsin kuningas Kaarle XI  
Museovirasto



Ruotsin kuningas Fredrik I  
Museovirasto

**Finto.fi**

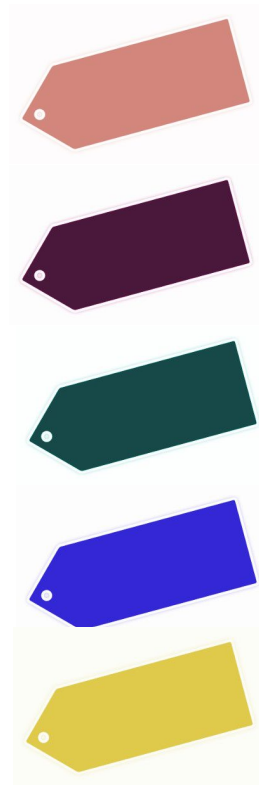
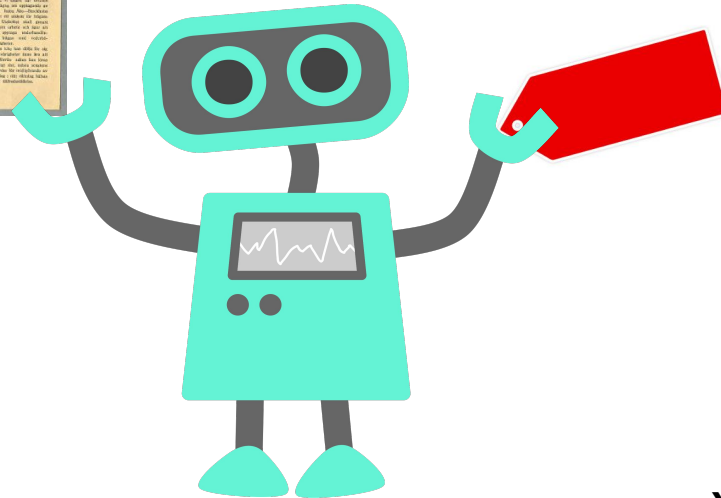
Where we publish thesauri, classifications, ontologies etc.  
for use by libraries, archives, museums, media, students...

**Subject indexing  
vocabularies:**

[General Finnish  
Ontology YSO](#)

[KOKO Ontology](#)

...and many more



YSO, Yleinen Suomalainen Ontologia  
with 30,000 subjects

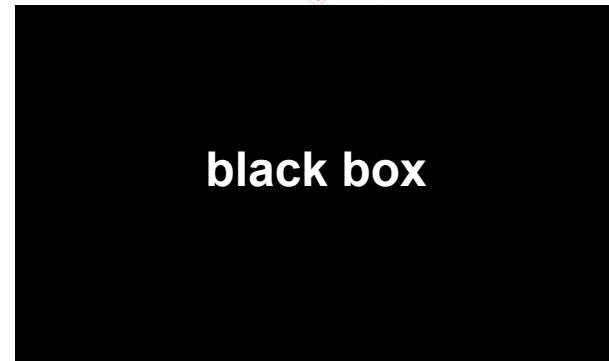
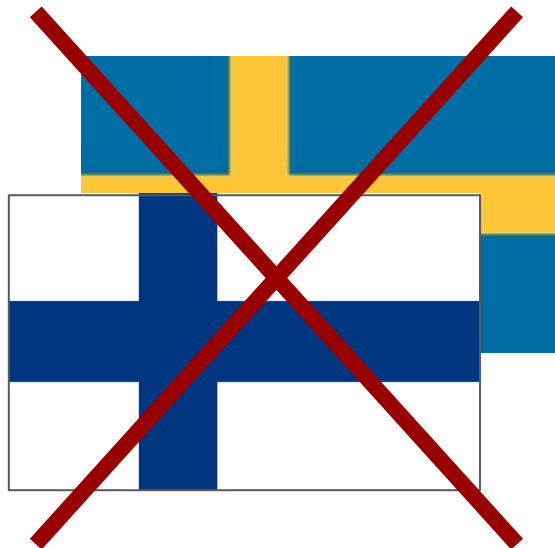


OPEN  
CALAIS



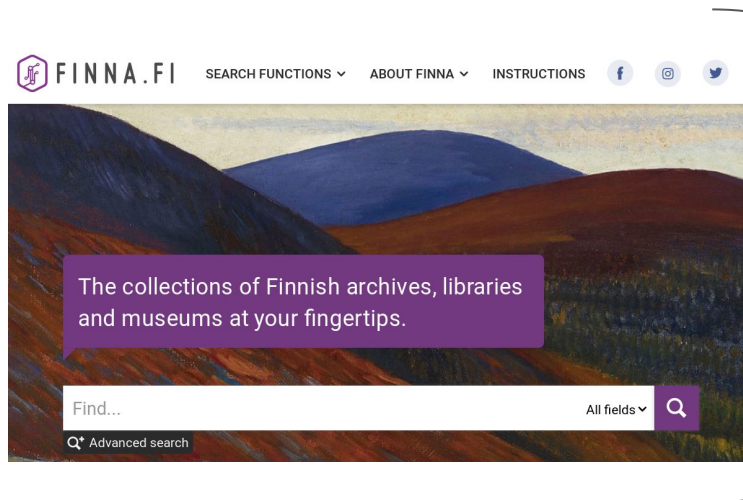
THOMSON REUTERS







# Machine learning using existing metadata



annif

## Lexical vs. Associative algorithms for subject indexing

## Lexical approaches: e.g. Maui

Match the **terms** in a document to **terms** in a controlled vocabulary

***“Renewable resources are a part of Earth's natural environment and the largest components of its ecosphere.”***

yso:p14146

“renewable natural resources”

**Associative** approaches (TFIDF, fastText, Omikuji ...)

Learn which **concepts** are correlated with which **terms** in documents, based on training data



# Algorithms used in Annif

- **TF-IDF similarity**

Baseline bag-of-words similarity measure. Implemented with the [Gensim](#) library.

- **[fastText](#)** by Facebook Research

Uses word embeddings (similar to [word2vec](#)) and resembles a neural network architecture.

- **[Vowpal Wabbit](#)**, originally by Yahoo! Research, now Microsoft Research

**Online machine learning** system, also suitable for multi-class and multi-label classification

- **[Parabel](#)** and **[Bonsai](#)**

**Tree-based algorithms** for extreme multi-label classification. Implemented with the [Omikuji](#) library.

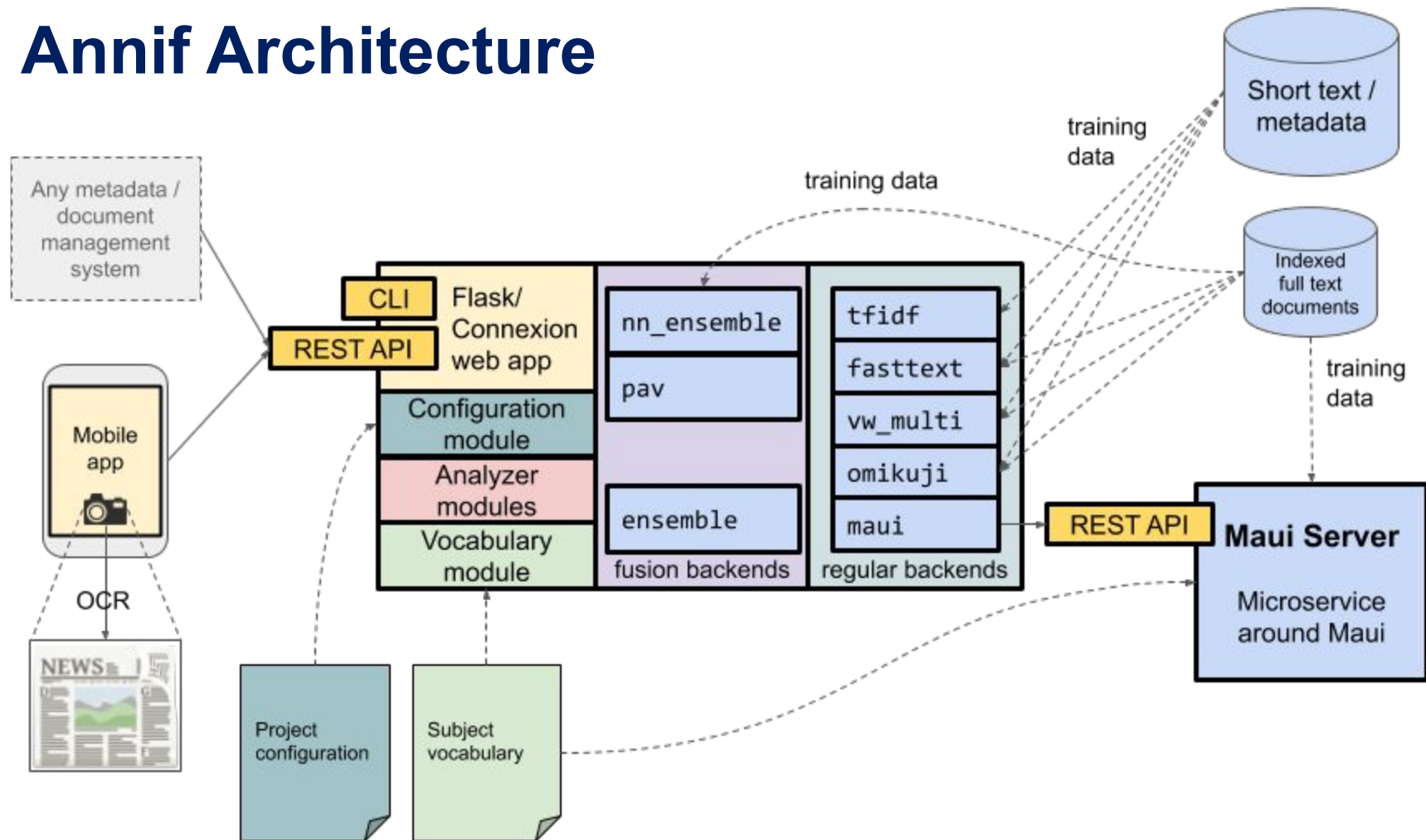
- **Maui** using MauiServer REST API

[MauiServer](#) is a microservice wrapper around the lexical [Maui](#) automated indexing tool.

Algorithms may be used **alone**, or in combinations, **ensembles**



# Annif Architecture



# Form for testing at [annif.org](https://annif.org)

## Try Annif!

Text to analyze:

How the humble potato changed the world

A staple food for cultures across the globe, the tuber has emerged as a nutritional giant and the friend of peasants, rulers and sages. Even today, its possibilities are endless.

In his 1957 essay collection *Mythologies*, the French philosopher and literary critic Roland Barthes called chips (*la frite*), a food that comes from a crop native to the Americas, “patriotic” and “the alimentary sign of Frenchness”.

Despite its origins in the Andes, it’s an incredibly successful global food

Just a century earlier, a potato disease prompted a famine that halved Ireland’s population in a few years, producing a decades-long cascading effect of social and economic turmoil. And as you read these lines, the world’s leading potato producers today are China, India, Russia and Ukraine, respectively.

**YSO model**  
trained on Finna data

Project (vocabulary and language):

YSO Ensemble English

Analyze

## Results



# Accessing Annif

**Command line interface** - setup and administration  
- training models  
- testing and evaluating models  
- bulk indexing of documents

**Web user interface** - interactive testing of models

**REST API** - integrating Annif services to other systems

# API access example

“The quick brown fox jumped over the lazy dog.”

suggest

annif

[api.annif.org](https://api.annif.org)

```
results=[  
  {uri="<http://www.yso.fi/onto/yso/p2228>", score=0.2595, label="red fox"},  
  {uri="<http://www.yso.fi/onto/yso/p5319>", score=0.2039, label="dog"},  
  {uri="<http://www.yso.fi/onto/yso/p8122>", score=0.1946, label="laziness"},  
  {uri="<http://www.yso.fi/onto/yso/p25726>", score=0.1285, label="brown"},  
  {uri="<http://www.yso.fi/onto/yso/p4760>", score=0.1220, label="triple jump"}  
]
```



NatLibFi / Annif

Unwatch 7 Unstar 23 Fork 3

Code Issues 20 Pull requests 0 Projects 0 Wiki Insights Settings

Annif is a multi-algorithm automated classification and subject indexing tool for libraries, archives and museums. This repository is used for developing a production version of the system, based on ideas from the initial prototype. <http://annif.org>

subject-indexing python machine-learning code4lib classification rest-api flask-application connexion Manage topics

766 commits 7 branches 48 releases 5 contributors View license

Branch: master New pull request Create new file Upload files Find file Clone or download

osma add Zenodo DOI badge Latest commit d832514 2 days ago

annif	refactor: split off JSON input to document corpus conversion in rest ...	2 days ago
tests	CLI unit test for trying to learn when backend doesn't support it	2 days ago
.codeclimate.yml	more comprehensive Code Climate configuration	a year ago
.codecov.yml	Codecov should ignore setup.py	10 months ago
.coveragerc	Generate Codecov reports	2 years ago
.gitignore	Add virtualenv (default? de-facto?) folder to gitignore	15 days ago
.lgtn.yml	Add LGTM configuration excluding fasttext	5 months ago
.scrutinizer.yml	Try to fix pipenv/pip compatibility issue pypa/pipenv#2924 within Scr...	5 months ago
.travis.yml	install deb packages using apt addon (even though they're unnecessary...	a month ago

# Annif on GitHub

Python 3.6+ code base

Apache License 2.0

Fully unit tested (99% coverage)

PEP8 style guide compliant

Usage [documentation](#) in the wiki

<https://github.com/NatLibFi/Annif>



[pypi.org/project/annif/](https://pypi.org/project/annif/)



[quay.io/natlibfi/annif](https://quay.io/natlibfi/annif)

# annif 0.47.1

`pip install annif`



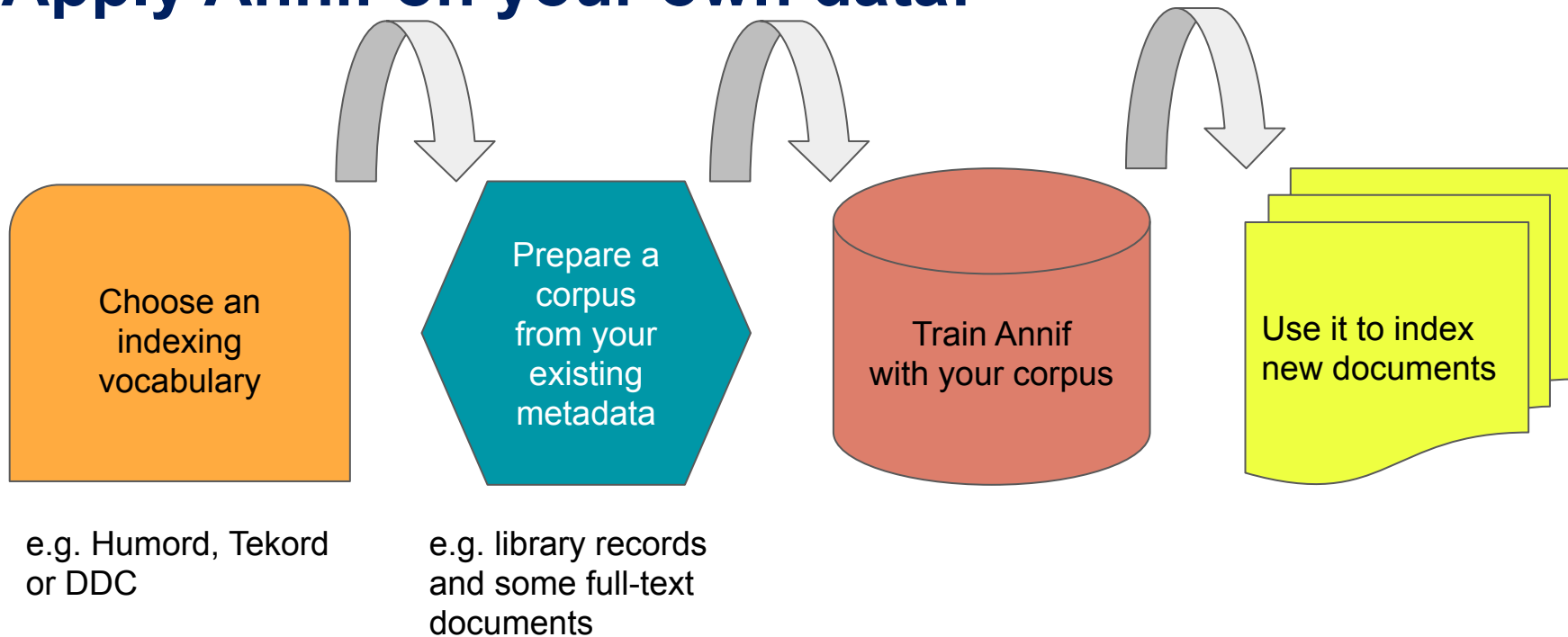
Automated subject indexing and classification tool

```
jmminkin@lx8-9811-008: /home/local/jmminkin/git/Annif
jmminkin@lx8-9811-008: /home/local/jmminkin/git/Annif 99x35
(Annif) jmminkin@lx8-9811-008: /home/local/jmminkin/git/Annif$ annif
Usage: annif [OPTIONS] COMMAND [ARGS]...

Options:
  --version  Show the flask version
  --help     Show this message and exit.

Commands:
  clear      Initialize the project to its original, untrained state.
  eval       Analyze documents and evaluate the result.
  index      Index a directory with documents, suggesting subjects for...
  learn      Further train an existing project on a collection of...
  list-projects  List available projects.
  loadvoc    Load a vocabulary for a project.
  optimize   Analyze documents, testing multiple limits and thresholds.
  routes     Show the routes for the app.
  run        Run a development server.
  shell      Run a shell in the app context.
  show-project  Show information about a project.
  suggest    Suggest subjects for a single document from standard input.
  train      Train a project on a collection of documents.
(Annif) jmminkin@lx8-9811-008: /home/local/jmminkin/git/Annif$
```

# Apply Annif on your own data!



**Annif used in production**

# JYX repository, University of Jyväskylä

Students upload their Master's and doctoral theses, Annif suggests subjects\*

## Keywords

<b>Keyword suggestions</b> <i>Choose valid keywords by clicking</i>	<ul style="list-style-type: none"><li><input type="checkbox"/> information management systems [YSO]</li><li><input type="checkbox"/> metadata [YSO]</li><li><input type="checkbox"/> connections (technical systems) [YSO]</li><li><input type="checkbox"/> content management [YSO]</li><li><input type="checkbox"/> multimedia (information technology) [YSO]</li><li><input type="checkbox"/> digital libraries [YSO]</li><li><input type="checkbox"/> XML [YSO]</li><li><input type="checkbox"/> semantic web [YSO]</li><li><input type="checkbox"/> open source code [YSO]</li><li><input type="checkbox"/> open data [YSO]</li><li><input type="checkbox"/> user-centeredness [YSO]</li><li><input type="checkbox"/> archives (memory organisations) [YSO]</li><li><input type="checkbox"/> seeking [YSO]</li><li><input type="checkbox"/> Works [YSO]</li><li><input type="checkbox"/> cloud services [YSO]</li><li><input type="checkbox"/> electronic publications [YSO]</li></ul>
<b>Your own keywords</b> <i>Comma separated list</i>	<input type="text" value="keyword 1, keyword 2"/>

Implemented using  
DSpace &  
[GLAMpipe](#)  
by Ari Häyrynen

\*from YSO =  
General Finnish  
Ontology

# Osuva repository, University of Vaasa

Same as JYX: students upload their Master's and doctoral theses, Annif suggests subjects

Pilot started  
2.3.2020,  
implementation  
by Anis  
Moubarik.

Asiasanat:

## Annif-ehdotukset

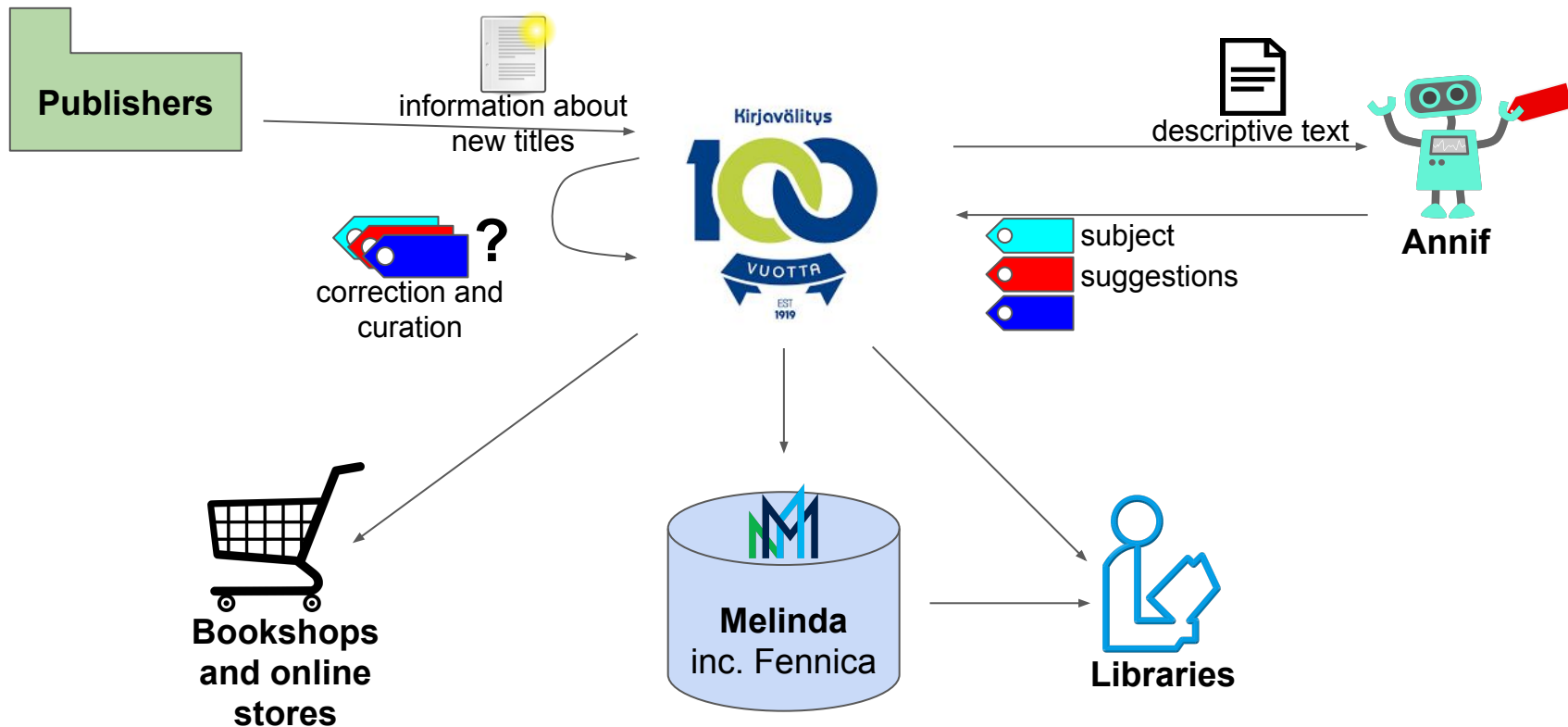
- |  |  |
|--|--|
| <input type="checkbox"/> working abroad      | <input type="checkbox"/> families (groups)       |
| <input type="checkbox"/> career development  | <input type="checkbox"/> managers and executives |
| <input type="checkbox"/> career              | <input type="checkbox"/> human resources         |
| <input type="checkbox"/> adaptation (change) | <input type="checkbox"/> work                    |
| <input type="checkbox"/> expatriates         | <input type="checkbox"/> returnees (immigrants)  |

Lisää

Lisää

Syötä asiasanat, jokainen asiasana omaan kenttäänsä. Paina siis jokaisen asiasanan jälkeen Lisää-nappia. Kirjoita tarvittava määrä asiasanan alkua, jolloin ennakoiva tekstinsyöttö ehdottaa asiasanoja. Muista myös valita yllä olevasta laatikosta Annif-ehdotukset, jotka perustuvat edellisessä vaiheessa syöttämäsi kokotekstin sisältöön.

# Kirjavälitys Oy - logistics company serving bookstores and libraries



# Finto AI - automated subject indexing tool and API service



About Feedback

[suomeksi](#) [på svenska](#)

Finto AI suggests subjects for a given text. It's based on Annif, a tool for automated subject indexing. [Read more...](#)

## API service

Finto AI is also an API service that can be integrated to other systems.

[Lisätietoja](#) | [OpenAPI-kuvaus](#)

### Enter text to be indexed

In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals. Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.[1] Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic human behavior, especially those that associate with the human mind, such as natural language processing.

As machines become increasingly capable, some researchers argue that they are often removed from the definition of intelligence. In 1950, Alan Turing proposed the Turing Test, in which a human judge asks questions of a human and a machine. The machine that best imitates human behavior wins. In 1956, the term "artificial intelligence" was coined at the Dartmouth Conference. The term "AI" is often used to describe machines (or computers) that mimic human behavior, especially those that associate with the human mind, such as natural language processing. The term "AI" is often used to describe machines (or computers) that mimic human behavior, especially those that associate with the human mind, such as natural language processing.

**Launched in  
May 2020**

[ai.finto.fi](https://ai.finto.fi)

### Subject indexing

Vocabulary and text language

YSO English

Maximum # of suggestions

10

15

20

Get subject suggestions

### Suggestions

- [artificial intelligence](#)
- [machine learning](#)
- [intelligence \(mental properties\)](#)
- [information technology](#)
- [computational science](#)
- [computer science](#)
- [computers](#)
- [computer-assisted teaching](#)
- [learning](#)
- [automation](#)



# Tagging at Yle

- All content tagged:  
subject/topic, but also  
e.g. genre and  
atmosphere

# Tagging at Yle

- All content tagged: subject/topic, but also e.g. genre and atmosphere

## Source ontologies / Vocabularies

### KOKO

- Collection of Finnish (general and special field) ontologies
- By The National Library of Finland / Finto service
- About 55 000 concepts

### Wikidata

- Free, collaborative, multilingual database. Structured data to provide support for e.g. Wikipedia
- About 87 million items

+ Freebase  
+ some internal vocabularies

Manual tagging

### Leiki

- By DoubleVerify
- About 150 000 concepts

Automatic tagging

# Tagging at Yle

- All content tagged: subject/topic, but also e.g. genre and atmosphere

## Content management systems

### Articles

Escenic

+ others

### Images

MAM Avid Interplay

+ others

### Programs and clips

Whats'On

Ceiton

MAM Avid Interplay

+ others

APIs for searching concepts, storing concepts and content-to-concept relations, and for analysing text

## Source ontologies / Vocabularies

### KOKO

- Collection of Finnish (general and special field) ontologies
- By The National Library of Finland / Finto service
- About 55 000 concepts

Manual tagging

### Wikidata

- Free, collaborative, multilingual database. Structured data to provide support for e.g. Wikipedia
- About 87 million items

+ Freebase  
+ some internal vocabularies

### Leiki

- By DoubleVerify
- About 150 000 concepts

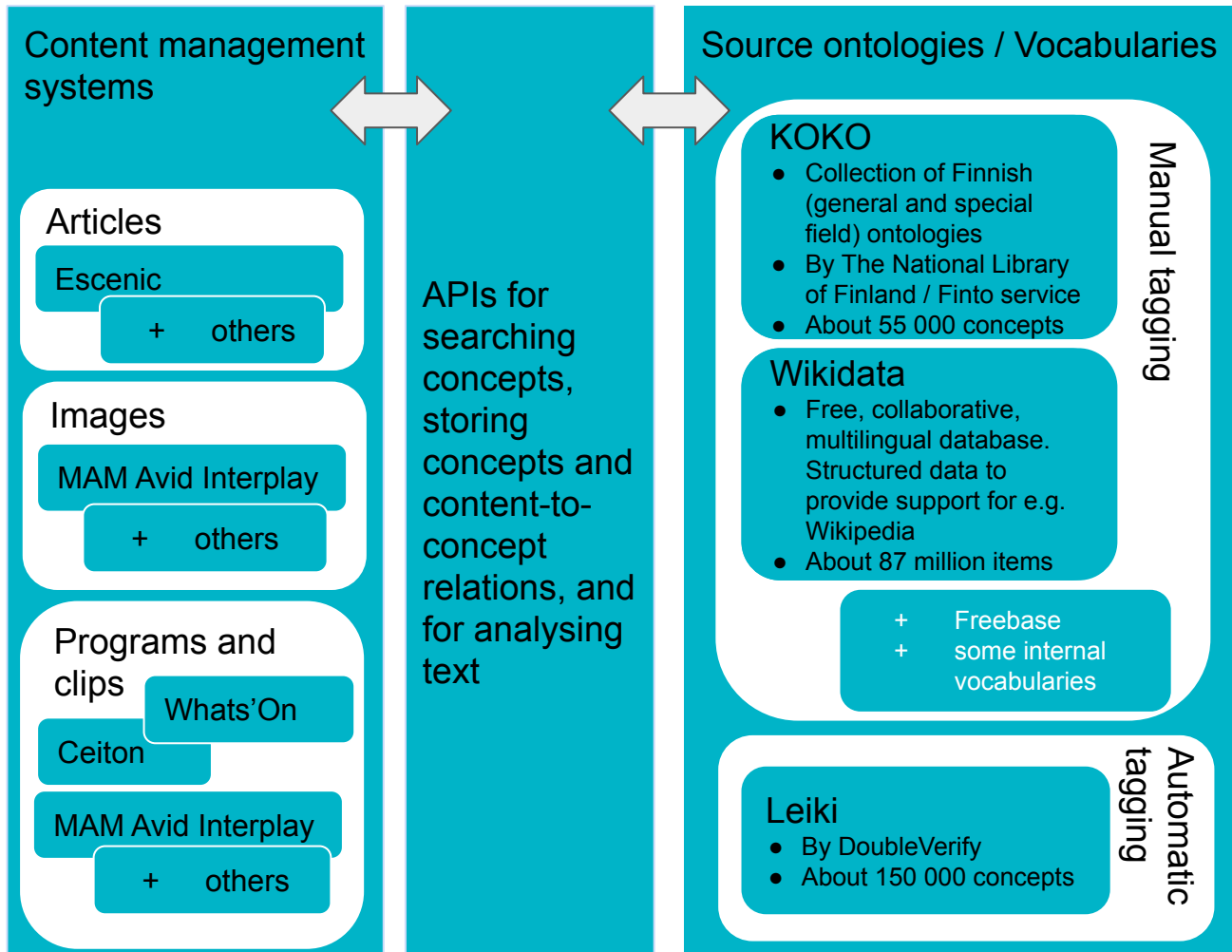
Automatic tagging

# Tagging at Yle

- All content tagged: subject/topic, but also e.g. genre and atmosphere

## Yle vocabulary

- = Concepts used for tagging of content at Yle
- Concepts from different source vocabularies
- About 200 000 concepts



# Automatic Tagging Currently at Yle

- Since 2015 by Leiki / DoubleVerify (<http://www.leiki.com/>)
- Articles in three languages: Finnish, Swedish, English
- Semi-automatic: 15 proposals for tags into CMS
  - > Validation
  - > Missing tags added manually
- The quality: Largely good
- The biggest challenge: Leiki's own ontology
  - > Content tagged with concepts from different ontologies are not linked in Yle's services
  - > Need to be mapped
- Request for information (RFI) in autumn 2019: No other service or tool available in Finland for our needs, only development projects

# Annif at Yle

## Why are we interested in Annif?

- To learn what kind of tagging we can produce with Yle Vocabulary
  - Online services and applications which use tags for e.g. personalisation and we have to ensure continuity
- To get experience of a method which can be further developed by ourselves and customized to suit our needs
- We have training data!
- Already the first experiments were promising!  
(Pekka Kauranen, autumn 2019)
- Part of “Yle Metadata Machinery” development
  - Combination of various methods and services for automatic metadata extraction

# Training Annif at Yle

- Annif backends used for training
  - Omikuji (AttentionXML)
  - Maui
  - nn\_ensemble (sources: Omikuji:2, Maui:1)
- Training text corpora
  - Articles, published 1998 - 18.1.2020
  - Short text document corpus: Titles and introductions of articles
    - Finnish: 1 048 756 articles
    - Swedish: 369 572 articles
    - Used to train Omikuji
  - Full text document corpus: Titles, introductions and bodies of articles
    - Used to train Maui, about 3000 articles / language
    - Used to train nn\_ensemble, about 500 articles / language
- Training vocabulary
  - Yle vocabulary (sources: KOKO, Wikidata, Leiki, Freebase)
  - Finnish: 171 979 tags
  - Swedish: 168 337 tags

# Training Annif at Yle

- Annif backends used for training
  - Omikuji (AttentionXML)
  - Maui
  - nn\_ensemble (sources: Omikuji:2, Maui:1)
- **Training text corpora**
  - Articles, published 1998 - 18.1.2020
  - Short text document corpus: Titles and introductions of articles
    - **Finnish: 1 048 756 articles**
    - **Swedish: 369 572 articles**
    - Used to train Omikuji
  - Full text document corpus: Titles, introductions and bodies of articles
    - Used to train Maui, about 3000 articles / language
    - Used to train nn\_ensemble, about 500 articles / language
- **Training vocabulary**
  - Yle vocabulary: Combination of tags from different source vocabularies (KOKO, Wikidata, Leiki, Freebase)
  - **Finnish: 171 979 tags**
  - **Swedish: 168 337 tags**



# Annif-Leiki Comparison

- One method to evaluate the quality of Annif:  
Annif and Leiki tagging compared by human evaluators
- About 100 Finnish and Swedish articles and their tags
  - Subject areas: business, science, culture, sport
- 28 Yle test persons
- Evaluation form (Google Form):
  - Article  
+ 10 most relevant tags from both tools/services in random order
  - Scale of relevance:  
Essential - ok - not relevant - wrong

Annif-Leiki-vertailu

Alkuperäiseen artikkeliin pääset tästä linkistä: <https://yle.fi/aihe/yle-id/20-299771>

Article

Planeettavaluutta yksi ratkaisu maapallon kestävyyskriisiin – suomalaistutkijoiden keinot ihmisten toiminnan muuttamiseksi  
Voisimme järjestää arkiympäristömme ja viljelytapamme paljon kestävämmällä tavalla. Raha voisi olla kriittisiin luonnonvaroihin sidottua planeettavaluuttaa, jolloin talouden

Arvioi asiasanojen osuvuutta

Voit katsoa asiasanan kuvauksen käymällä osoitteessa "[https://meta.api.yle.fi/v1/concepts.json?app\\_id=HIEKKALAATIKKO&app\\_key=HIEKKALAATIKKO&yle\\_id=YLE\\_ID](https://meta.api.yle.fi/v1/concepts.json?app_id=HIEKKALAATIKKO&app_key=HIEKKALAATIKKO&yle_id=YLE_ID)" ottamalla YLE\_ID:n asiasanan suluista

Evaluation scale

	essential keskeinen	ok ok	non relevant epärelevantti	wrong väärä
professorit (18-211223)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
aurinkokunnat (18-8786)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Markku Wilenius (18-295234)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
fossiiliset polttoaineet (18-2396)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
ilmasto (18-215534)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
tulevaisuus (18-211385)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Tags

# Comparison: Overall Results in Subject Areas

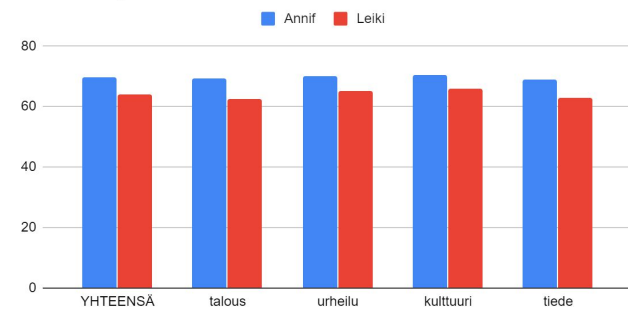
## Finnish

Annif performed slightly better than Leiki  
= more essential + ok,  
less not relevant + wrong tags

Culture: The only subject area where Leiki performed slightly better: more ok, less wrong tags

Essential + ok (% of all tags)  
TOTAL - business - sport - culture - science

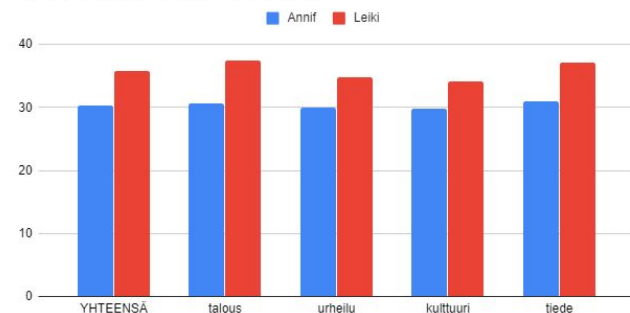
Keskeiset ja OK-asiasanat



Suomi: Keskeisiksi arvioitut ja OK-asiasanat aihealueittain (% kaikista arvioituista asiasanoista)

Not relevant + wrong (% of all tags)  
TOTAL - business - sport - culture - science

Epärelevantit ja väärät asiasanat

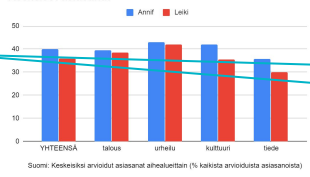


Suomi: Epärelevantteiksi ja vääräksi arvioitut asiasanat aihealueittain (% kaikista arvioituista asiasanoista)

essential

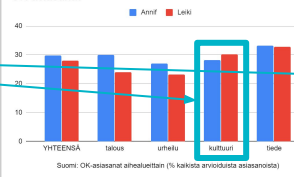
ok

Keskeiset asiasanat



Suomi: Keskeisiksi arvioitut asiasanat aihealueittain (% kaikista arvioituista asiasanoista)

OK-asiasanat



Suomi: OK-asiasanat aihealueittain (% kaikista arvioituista asiasanoista)

non relevant

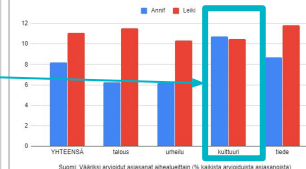
wrong

Epärelevantit asiasanat



Suomi: Epärelevantteiksi arvioitut asiasanat aihealueittain (% kaikista arvioituista asiasanoista)

Väärät asiasanat



Suomi: Vääräksi arvioitut asiasanat aihealueittain (% kaikista arvioituista asiasanoista)

# Comparison: Overall Results in Subject Areas Swedish

Annif performed better than Leiki in all subject areas = more essential + ok, less not relevant + wrong tags

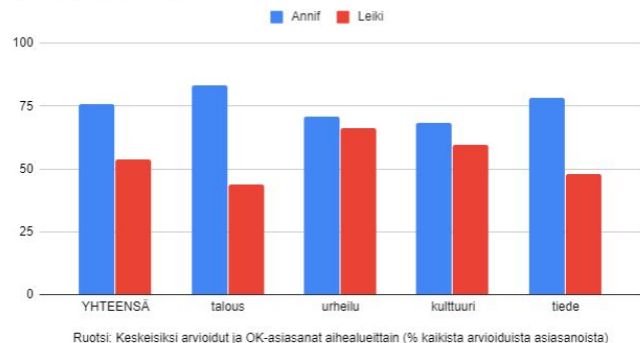
Differences bigger than in Finnish

Biggest differences in business and science

Reasons?

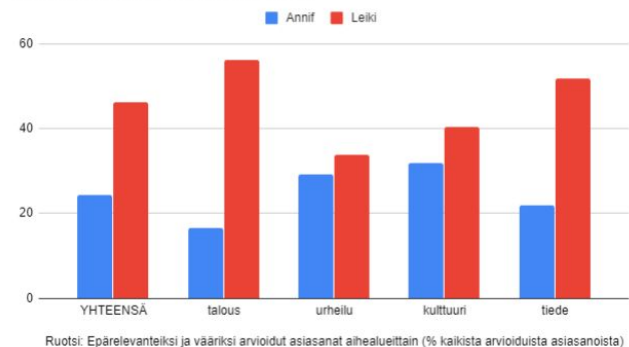
Essential + ok (% of all tags)  
TOTAL - business - sport - culture - science

Keskeiset ja OK-asiasanat



Not relevant + wrong (% of all tags)  
TOTAL - business - sport - culture - science

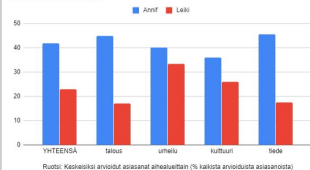
Epärelevantit ja väärät asiasanat



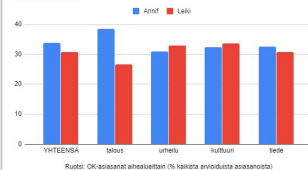
essential

ok

Keskeiset asiasanat



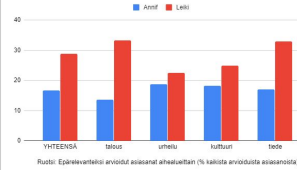
OK-asiasanat



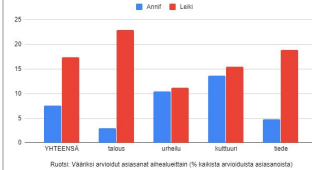
non relevant

wrong

Epärelevantit asiasanat



Väärät asiasanat



# Next Steps with Annif at Yle

- The results of the comparison encourage us to carry on testing Annif!
- Currently optimization of algorithms and vocabulary
- Test use in CMS to get more data
- Tests with other text types (to get automatic tagging for tv and radio programs):
  - Speech to text results
  - Program descriptions
  - Subtitles

# Thank you!

**annif**

Annif team:

Juho Inkinen



Mona Lehtinen



Osma Suominen



e-mail:  
firstname.lastname@helsinki.fi

Pekka Kauranen



yle

Ahti Ahde



**QVIK**

Pia Virtanen



yle

e-mail:  
firstname.lastname@yle.fi  
firstname.lastname@qvik.fi

# Questions?

## Topics for discussion, e.g.

- The quality of subject indexing / tagging:  
How to measure it and give feedback to the librarian / content producer?
- Using open source software and developing it yourselves  
vs. buying software as a service

# More about

Website: [annif.org](https://annif.org)

API: [api.annif.org](https://api.annif.org)

Source code: <https://github.com/NatLibFi/Annif>

Wikipages: <https://github.com/NatLibFi/Annif/wiki>

User group: [annif-users@googlegroups.com](mailto:annif-users@googlegroups.com)

Tutorial: <https://github.com/NatLibFi/Annif-tutorial>

Article: Suominen, O., 2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, 29(1), pp.1–25. DOI: <http://doi.org/10.18352/lq.10285>

# More about tagging at Yle

Pia Virtanen: Tagging Content at Yle. Finnish high school matriculation exams and other use cases for tags.

Presentation at EBU MDN Workshop, 7.6.2018:

<https://docs.google.com/presentation/d/1neH0nB3PhZVLY2baGC4LJAPdiXsmwcSQqvNnqSatsEE>

Pia Virtanen: Tagging Content at the Finnish Broadcasting Company Yle.

Presentation at EBU MDN Workshop, 7.6.2016:

[https://tech.ebu.ch/docs/events/mdn2016/presentations/Pia\\_Virtanen\\_YLE\\_EBU%20MDN%202016.pdf](https://tech.ebu.ch/docs/events/mdn2016/presentations/Pia_Virtanen_YLE_EBU%20MDN%202016.pdf)

Pia Virtanen, Kim Viljanen, Mikael Hindsberg:  
YLE's Meta-API: Improving the Findability of Web Content with Semantic Tagging. In: Tech Report 019. EBU-MIM Semantic Web Activity Report, 2015. Annex 9, p. 43-56.

<https://tech.ebu.ch/publications/tr019>

# Bio: Osma Suominen

Osma Suominen works as an information systems specialist at the National Library of Finland. He is currently working on automated subject indexing and the publishing of bibliographic data, including the Finnish national bibliography Fennica, as Linked Data. He is one of the creators of the Finto.fi thesaurus and ontology service and is leading development of the Skosmos vocabulary browser used in Finto. Osma Suominen earned his doctoral degree at Aalto University while doing research on semantic portals and quality of controlled vocabularies within the FinnONTO series of projects.

GitHub: [@osma](#)

Twitter: [@OsmaSuominen](#)

# Bio: Pia Virtanen

Pia Virtanen works as producer / metadata specialist at the Finnish Broadcasting Company Yle developing methods and practises of describing content, especially tagging of online content. She is responsible for leading the technical development needed for company-wide solutions in tagging, for data maintenance of the common “Yle vocabulary” as well as for supporting and instructing journalists in their tagging processes. Pia Virtanen is a trained translator and information specialist / librarian and working since 2005 at Yle.

*Please, be in contact,  
also e.g. if you would like to discuss vocabularies in general or vocabularies for  
tagging genre and atmosphere  
of tv and radio programs in particular!*

[pia.virtanen@yle.fi](mailto:pia.virtanen@yle.fi)



# Abstract of this presentation

Yle meets Annif - an open source tool for automated subject indexing

In the first part of our presentation we will introduce Annif, an open source tool for automated subject indexing. Annif is based on natural language processing and machine learning technologies. It can suggest subjects for documents using any subject vocabulary, in a variety of languages. Annif is being developed at the National Library of Finland and many cultural heritage institutions are starting to use it to support their subject indexing processes.

In the second part we will present a project where Yle trained Annif using over a million Finnish and Swedish language articles along with their subjects, drawn from a vocabulary of 170,000 concepts. We also evaluated the quality of Annif by comparing its results to the service we currently use for subject indexing of articles. Because the quality of Annif has been promising the project will be carried on. The project made us think about the quality of subject indexing and how it can be measured.

Osma Suominen – National Library of Finland, Pia Virtanen – Yle