



Annif and Finto AI: DIY automated subject indexing from prototype to production

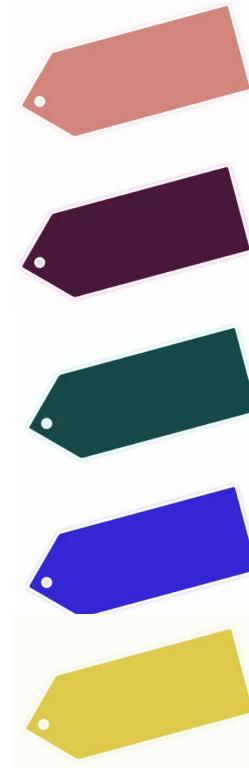
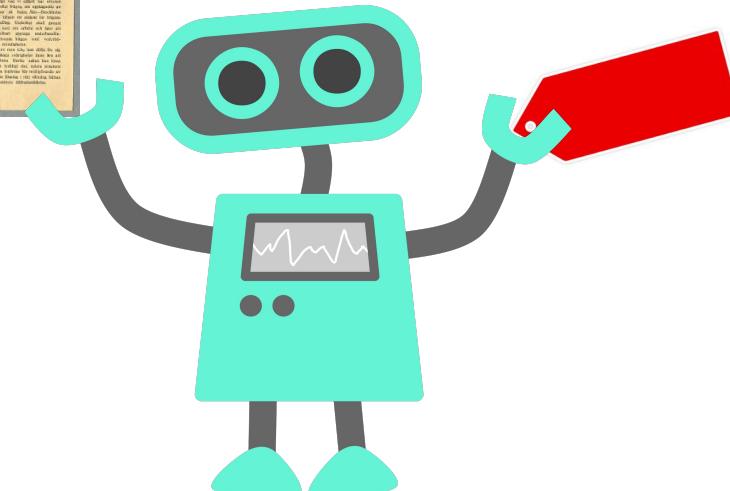
Osma Suominen, Mona Lehtinen, Juho Inkkinen
SWIB20, 23 November 2020



Outline

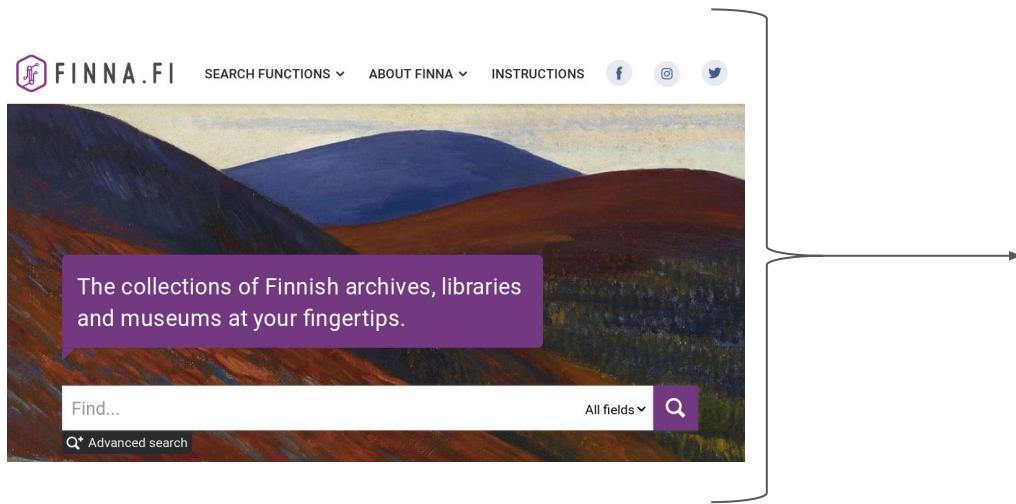
1. Development of Annif
2. Quality of automated subject indexing
3. Community building
4. Annif deployments
5. Lessons learned

1. Development of Annif



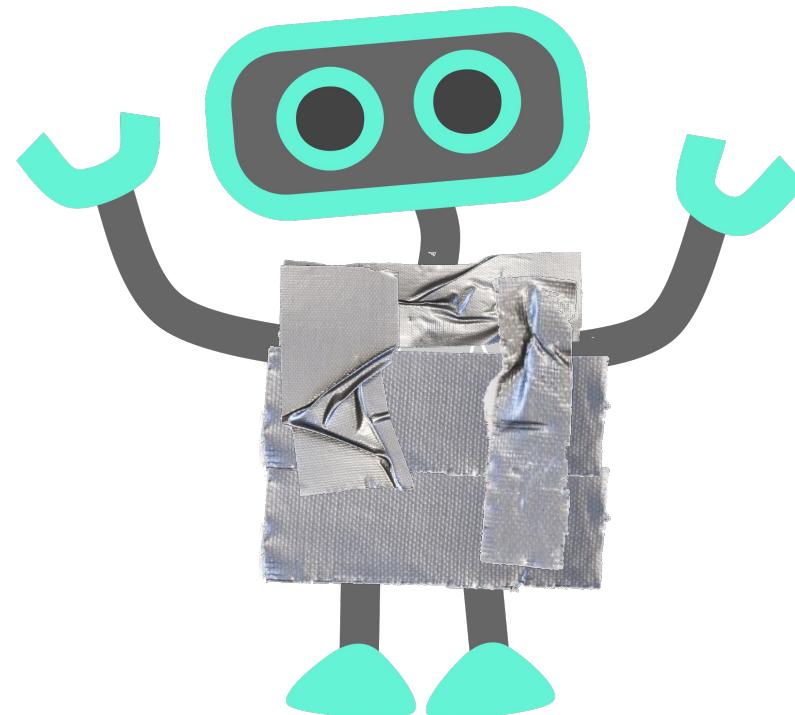
YSO, General Finnish Ontology
with 30,000+ subjects

Machine learning using existing metadata



annif

Early prototype (2017) got people excited



Starting points for Annif implementation (2018 →)

1. multilingual
2. independent of indexing vocabulary
3. support different subject indexing algorithms
4. CLI, Web user interface and REST API
5. community-oriented open source



[NatLibFi / Annif](#)

Code Issues Pull requests Projects 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

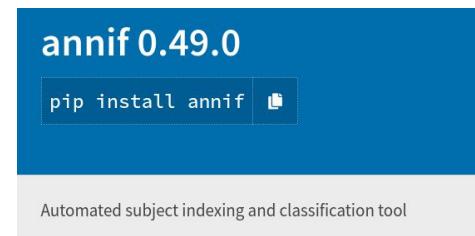
osma	add Zenodo DOI badge	Latest commit d832514 2 days ago
annif	refactor: split off JSON input to document corpus conversion in rest ...	2 days ago
tests	CLI unit test for trying to learn when backend doesn't support it	2 days ago
.codeclimate.yml	more comprehensive Code Climate configuration	a year ago
.codecov.yml	Codecov should ignore setup.py	10 months ago
.coveragerc	GenerateCodecov reports	2 years ago
.gitignore	Add virtualenv (default? de-facto?) folder to gitignore	15 days ago
.lgtm.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

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



Python package on PyPI



Docker images on Quay.io

2. Quality of automated subject indexing

Document collections for training and evaluation

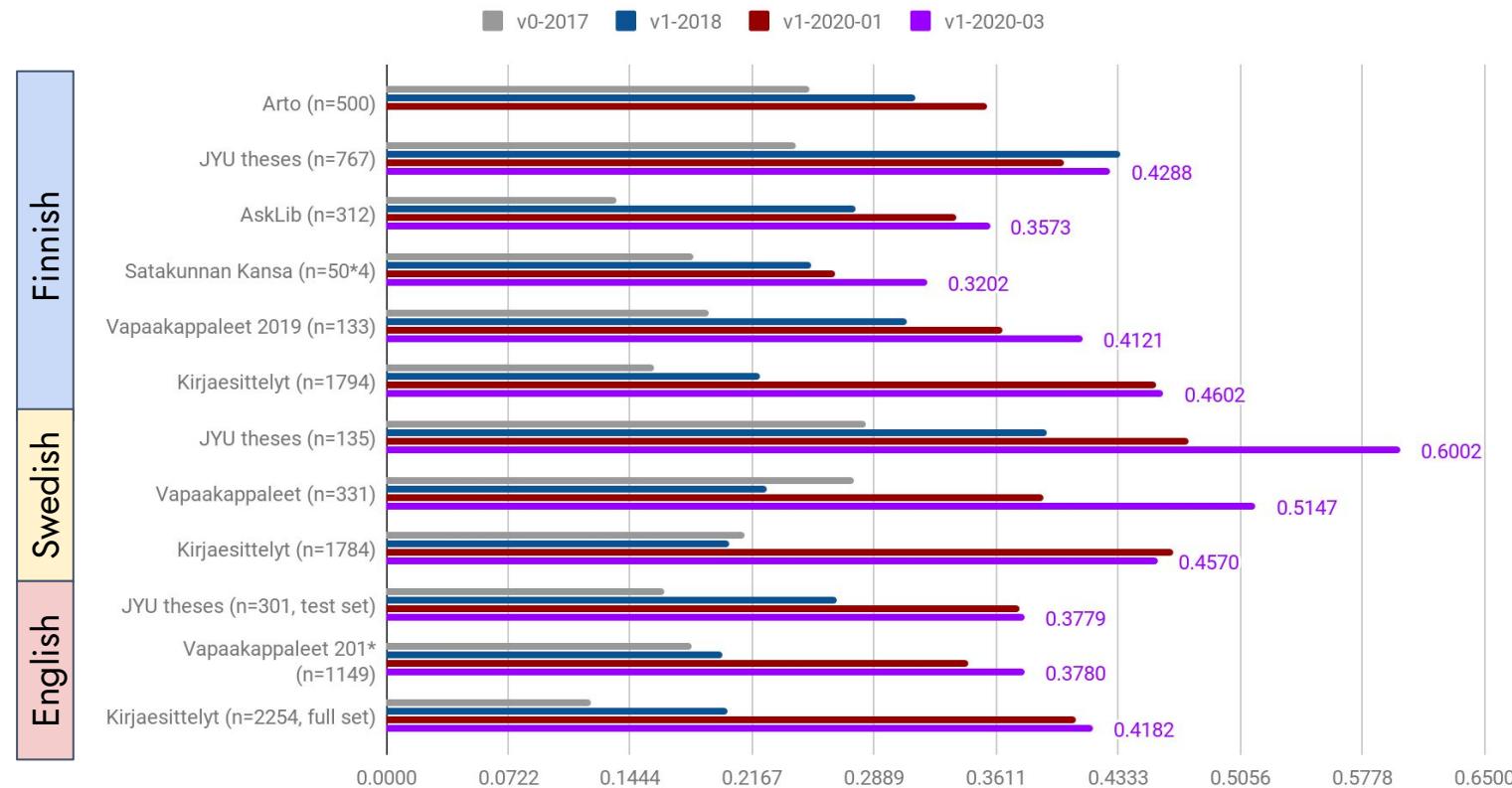
1. Metadata records from Finna.fi discovery system
2. Ask a Librarian question-answer pairs
3. Master's and Doctoral theses from University of Jyväskylä
4. Book descriptions from publishers (via Kirjavälitys Oy)
5. E-books from our electronic deposit system
6. ...

Converted to Annif corpus format & split into train/validate/test subsets

The ones we could republish are in the [Annif-corpora](#) repository GitHub

Comparison to “gold standard”

F1@5 scores for different test corpora and Annif API/model versions



Assessment by evaluators

At a workshop in 2019, **48 evaluators** evaluated subjects for **50 documents**. Subjects were given by either human indexers or four different algorithms.

The best ensemble algorithm (red bars) was not quite on the level of human indexers in quality scores (left), and significantly more of its suggestions were rejected (right).

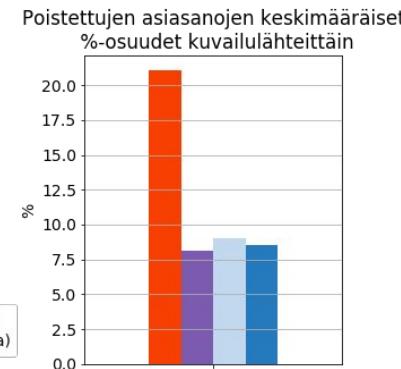
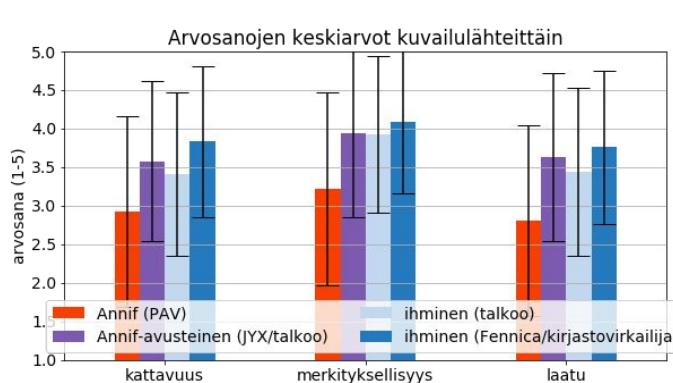


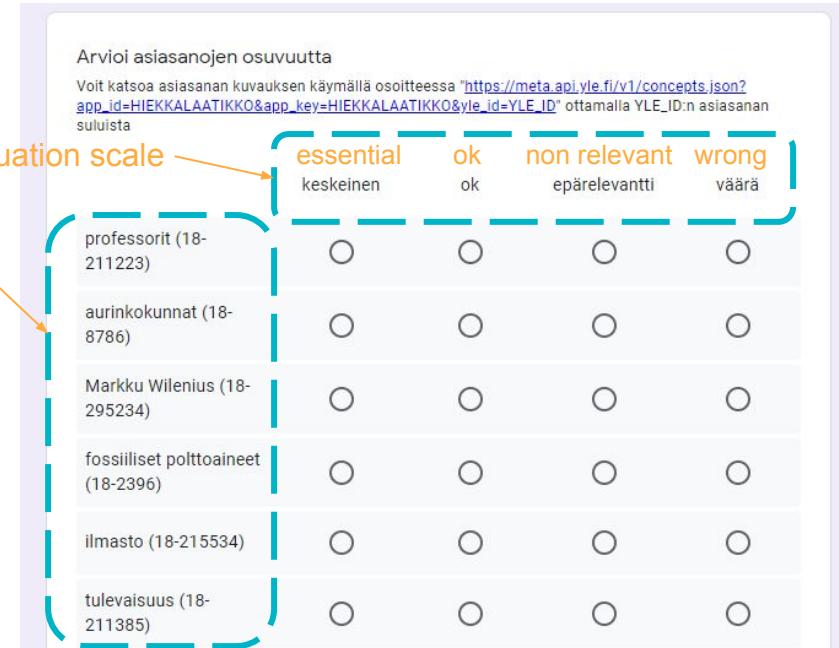
Photo: Mikko Lappalainen.

Annif-Leiki Comparison at Finnish Broadcasting Company Yle

- Annif vs Leiki (commercial service) tagging compared by 28 human evaluators at Yle
- About 100 Finnish and Swedish articles and their tags
 - business, science, culture, sport

Finnish: Annif **slightly better** than Leiki

Swedish: Annif **substantially** better than Leiki

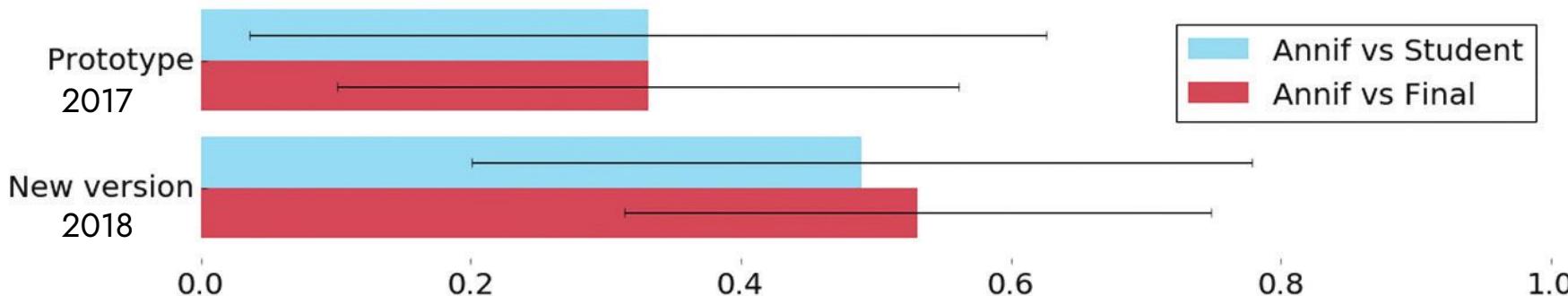


Evaluating in the context of an indexing workflow

JYX repository, University of Jyväskylä:

F1 similarity between Annif suggestions and the subjects

- a) chosen by the student (blue)
- b) confirmed by the JYX librarian (red)



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>

3. Community building

Web site with form for testing at annif.org

INPUT TEXT

SWIB focuses on Linked Open Data (LOD) in libraries and related organizations. It is well established as an event where IT staff, developers, librarians, and researchers from over the world meet and mingle and learn from each other. The topics of talks and workshops at SWIB revolve around opening data, linking data and creating tools and software for LOD production scenarios. These areas of focus are supplemented by presentations of research projects in applied sciences, industry applications, and LOD activities in other areas.

As usual, SWIB20 will be organized by the ZBW - Leibniz Information Centre for Economics and the North Rhine-Westphalian Library Service Centre (hbz). The conference language is English.

Would you like to share your experiences working on an interesting service, research topic or project – not just what you did, but also how you did it?

For this SWIB rendition we adjusted the formats to the online environment:

Presentations (15 minutes plus 5 q&a)

Practical workshops or tutorials (maximum 120 min)

PROJECT (VOCABULARY AND LANGUAGE)

YSO NN Ensemble English ▾

MAX # OF SUGGESTIONS

10 15 20

Get suggestions →



SUGGESTED SUBJECTS

- [data acquisition](#)
- [libraries](#)
- [linked open data](#)
- [information retrieval](#)
- [metadata](#)
- [data mining](#)
- [Finland](#)
- [information and communications technology](#)
- [information technology](#)
- [social media](#)

Wiki documentation on GitHub

- issues
- pull requests

Welcome to the Annif wiki!

- [Getting started](#)
- [System requirements](#)
- [Optional features and dependencies](#)
- [Usage with Docker](#)
- [Architecture](#)
- [Commands](#)
- [Web user interface](#)
- [Corpus formats](#)
 - Document corpus formats
 - Subject vocabulary formats
- [Project configuration](#)
- [Analyzers](#)
- [Achieving good results](#)
- [Reusing preprocessed training data](#)
- [Running as a WSGI service](#)
- [Backends/Algorithms supported by Annif](#)
 - Regular backends for automated subject indexing and classification
 - [Backend: TF-IDF](#)
 - [Backend: fastText](#)
 - [Backend: Omikuji](#)
 - [Backend: Maui](#)
 - [Backend: vw_multi](#)
 - Fusion/Ensemble backends that combine results from other backends
 - [Backend: Ensemble](#)

Welcome to the [Annif](#) users' mailing list / web forum! This list can be used for

- general discussion about Annif, its features and usage scenarios
- asking for help with installing or running Annif
- future directions for Annif
- announcements for new versions and other Annif-related news

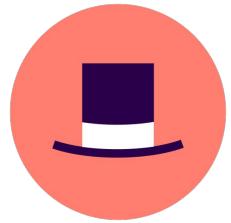
annif-users forum on Google Groups



	Remi Malessa	2	Loading OCLC FAST Vocabulary	— Actually, my server has around 32MB of memory and it s	14.55	
	osma.s..... , sale...@g...	4	Annif presentation and workshop at SWIB20 online conference	— Thanks Osma for sharing	10. marrask.	
	mona.l...@..., ... Annif...	13	Annif tutorial at the 2020 DCMI Virtual Conference	— Hi All, There is also a second online h	7. lokak.	
	sara.v...@kb.nl		Paper on Annif for categorizing laws	— Dear all, Some of you may have read my blogpost o	8. syysk.	
	haig...@..., osma.s.....	2	Running Annif as a WSGI service - connexion module error	— Hi Thomas! It looks like the P	7. syysk.	
	juho.k..... , osma.s.....	4	Results with Annif in employment service context	— Hi Juho! juho.k...@gmail.com kirjoitti	21. lokaku.	
	stephane5... , osma.s....	2	Matter of vocabulary	— Hi Stephane, Since you have a SKOS vocabulary, you can just load it	26. elok.	
	osma.s...@helsinki.fi		ANN: Annif 0.49 released	— Annif 0.49 has been released! https://github.com/NatLibFi/Annif/releases/tag/v0.49.0	30. heinäk.	

Hands-on Annif tutorial

for those who want to use Annif on their own



SWIB19
Semantic Web in Libraries

DCMI Virtual, 2020
September 14th-25th, 2020

SWIB20
Semantic Web in Libraries

The screenshot shows the 'annif tutorial' YouTube channel. A video thumbnail for 'Data sets & TFIDF tutorial' is displayed, featuring the ZBW logo at the bottom. To the right, a snippet of the 'Exercise 2: Set up and train a TFIDF project' page is shown, containing configuration details and code snippets.

Videos and exercises freely
available on YouTube & GitHub!



Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

4. Annif deployments

JYX repository, University of Jyväskylä

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

Keywords

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

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

*from YSO =
General Finnish
Ontology

Osuva repository, University of Vaasa

Trepo repository, University of Tampere

Theseus repository, Finnish universities of applied sciences

Same idea as JYX: students upload their theses,
Annif suggests subjects

Pilot started with Osuva in March
2020, others followed later.

DSpace extension implemented
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.

Finto AI - automated subject indexing tool and API service

fintoai

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 are capable of performing tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and language translation.

As machines become increasingly capable, computer programs may be able to perform more tasks without human intervention or supervision. Some AI programs are trained via machine learning, while others are explicitly programmed. The Church-Turing thesis says "AI is whatever we choose it to be".^[2] In general, the term "AI" is frequently excluded from things considered to be AI,^[3] having become a routine technology.^[6] Modern machine capabilities generally classified as AI include successfully understanding human speech,^[7] competing at the highest level in strategic game systems (such as chess and Go),^[8] autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

Launched in May 2020

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

Subject indexing for electronic deposits

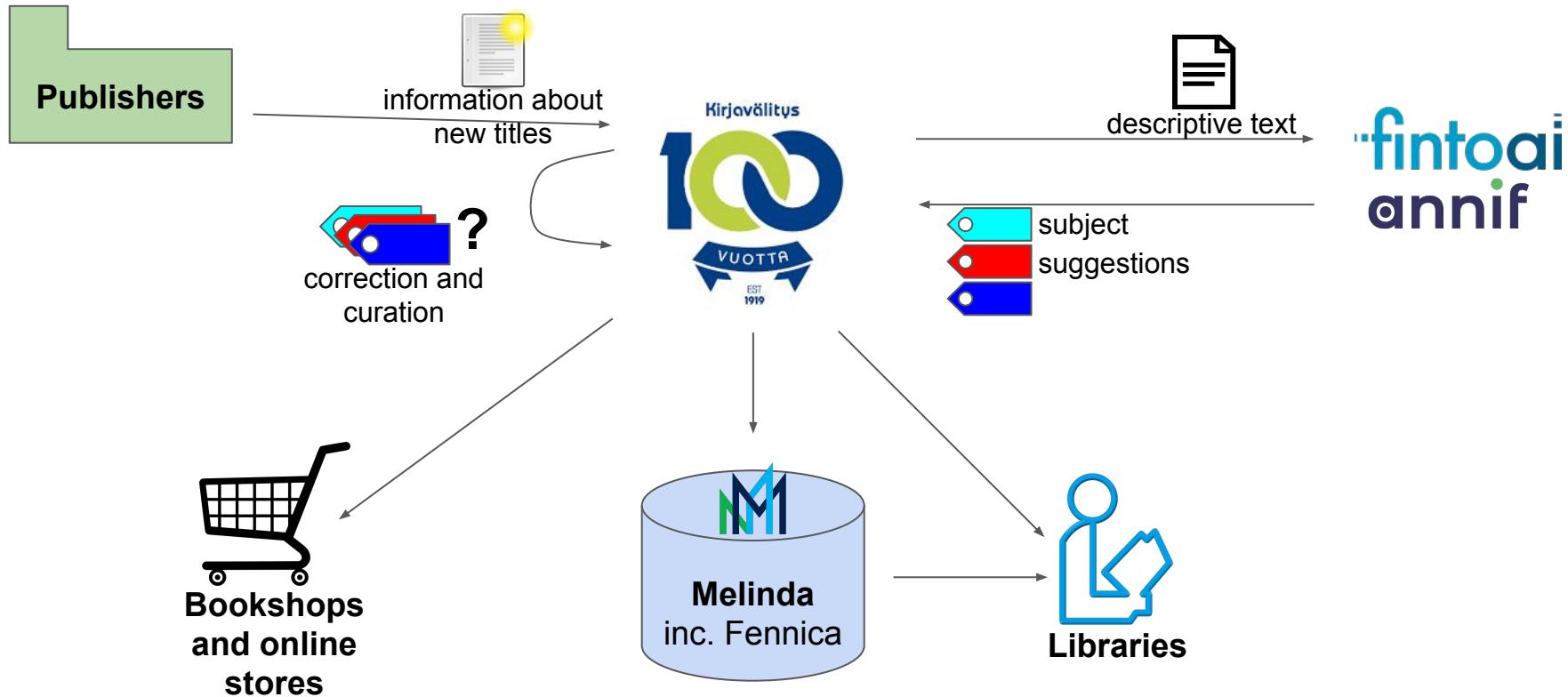
In November 2020, the National Library of Finland started using **Finto AI** to suggest subjects when processing electronic deposits submitted through the individual submission form.

The screenshot displays a web-based form for archiving digital publications. At the top right, there are language selection buttons for fi, sv, and en. The main header reads "Digitaalisten julkaisujen arkistointi". On the left, the National Library of Finland logo is visible. The form is divided into two main sections:

- Luovuttajan tiedot** (Donor information):
 - Fields marked with an asterisk (*) are mandatory: Yhteyshenkilö*, Sähköpostiosoitte*, Puhelinnumero*, Organisaatio.
 - A checkbox labeled "Muista nämä tiedot" (Remember these details) is present.
 - To the right, a note states "Tämä on pakollinen kenttä." (This is a mandatory field).
- Julkaisun tiedot** (Publication information):
 - Julkaisujen lukumäärä**: Radio buttons for "Luovutan yhden julkaisun" (I am handing over one publication) and "Luovutan useita julkaisuja" (I am handing over several publications).
 - Julkaisun tyyppi**: Radio buttons for "kirja" (book), "muotti" (sound recording), "äänite" (audiovisual work), and "muu" (other).
 - Perustiedot**: A field containing the text "ISBN (vivoilla)" (ISBN (available)).

Implementation: Erik Lindgren,
Mikko Merioksa, Satu Niininen

Kirjavälitys Oy - logistics company serving bookstores and libraries

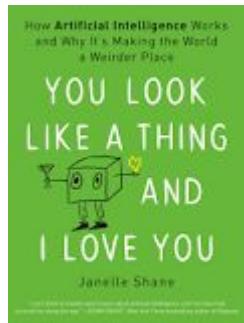


5. Lessons learned

Subject indexing is hard.

Humans have different perspectives and make understandable mistakes.

Algorithms make very silly mistakes.



Case in point:

Image recognition algorithms will frequently identify giraffes in pictures where there are none.

(Janelle Shane: You Look Like a Thing and I Love You)

Algorithms may be used **alone**, or in combinations, **ensembles**
Ensembles are nearly always better than individual algorithms



Lessons from evaluation

- The different evaluation approaches are complementary. (see Golub et al., 2016)
Not a good idea to look at just a single measure.
- Continuous and elusive process: it never stops...

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Hiom, D., and Lykke, M. 2016. A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, 67(1): 3-16.

Start by experimentation, move slowly towards production



image credit: @kettutatinukkeilee

With an API service such as Finto AI, implementing semi-automated indexing becomes easy; explaining it to users can be more challenging

Keywords

Keyword suggestions

Choose valid keywords by clicking

- information management systems [YSO]
- metadata [YSO]
- connections (technical systems) [YSO]
- content management [YSO]
- multimedia (information technology) [YSO]
- digital libraries [YSO]
- XML [YSO]
- semantic web [YSO]
- open source code [YSO]
- open data [YSO]
- user-centeredness [YSO]
- archives (memory organisations) [YSO]
- seeking [YSO]
- Works [YSO]
- cloud services [YSO]
- electronic publications [YSO]

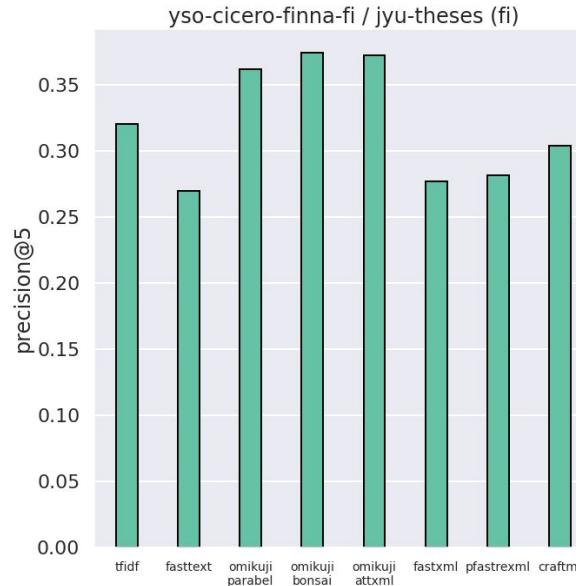
Your own keywords

Comma separated list

keyword 1, keyword 2



Collaboration is valuable! (1)



CSC has tested many state of the art text classification algorithms for us.
They discovered Omikuji, which is by far the best individual algorithm in Annif currently.

[High-Performance Digitisation](#) project 2018-2020, funded by INEA

Collaboration is valuable! (2)



[3]

KB } national library
of the netherlands



Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

[1] Martijn Kleppe, Sara Veldhoen, Meta van der Waal-Gentenaar, Brigitte den Oudsten, & Dorien Haagsma. (2019). Exploration possibilities Automated Generation of Metadata. DOI: <http://doi.org/10.5281/zenodo.3375192>

[2] Romein, C.A., Gruijter, M.D., & Veldhoen, S. (2020). The Datafication of Early Modern Ordinances. DH Benelux Journal, issue 2, 2020. <https://journal.dhbenelux.org/journal/issues/002/article-23-romein/article-23-romein.html>

[3] Lehtonen, T., Piukkula, J. Automaattinen asiasanoitus Radio- ja televisio-ohjelmaketokanta Ritvassa. Informaatiotutkimus 39 (1), 2020. DOI: <https://doi.org/10.23978/inf.88107>

[1,2]

Thank you!



Juho Inkinen



Mona Lehtinen



Osma Suominen

annif.org

These slides: <https://tinyurl.com/annif-swib20>