annif

fintoai

Annif and Finto AI:

DIY automated subject indexing from prototype to production

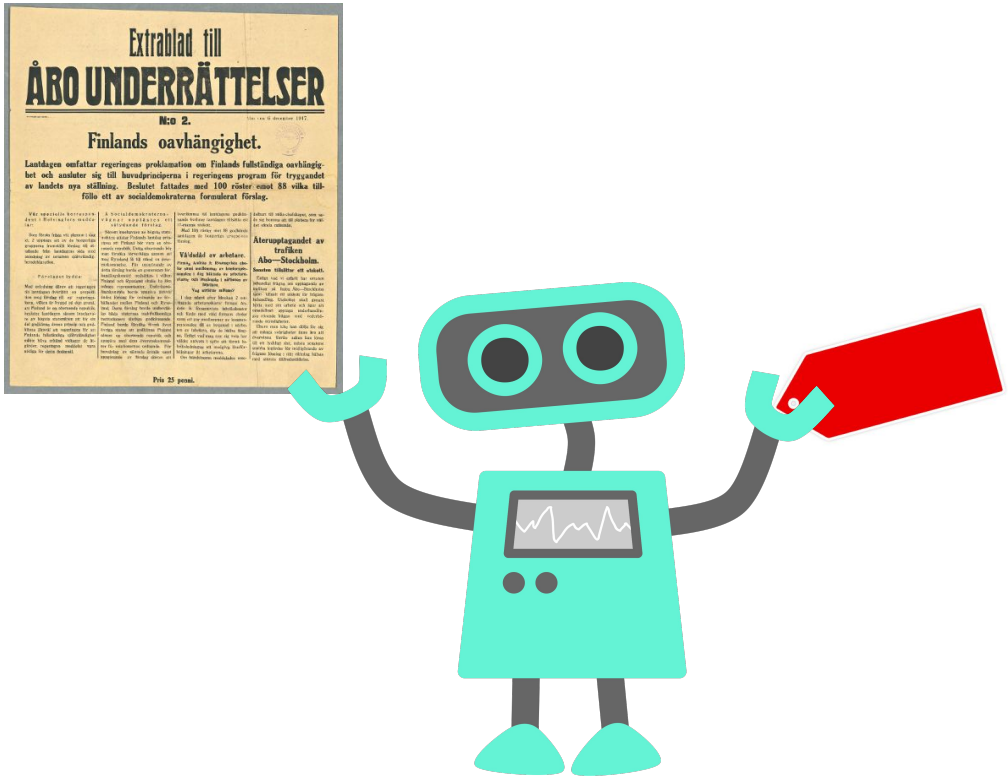
Osma Suominen, Mona Lehtinen, Juho Inkinen

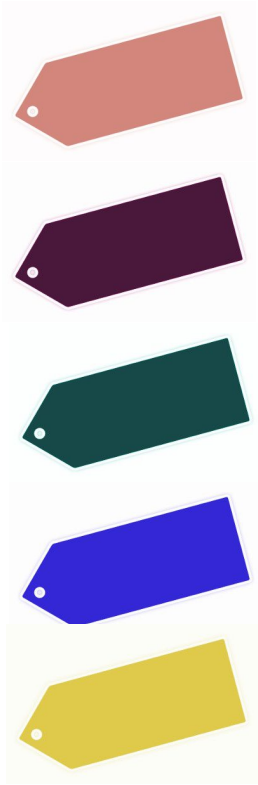
SWIB20, 23 November 2020

Outline

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

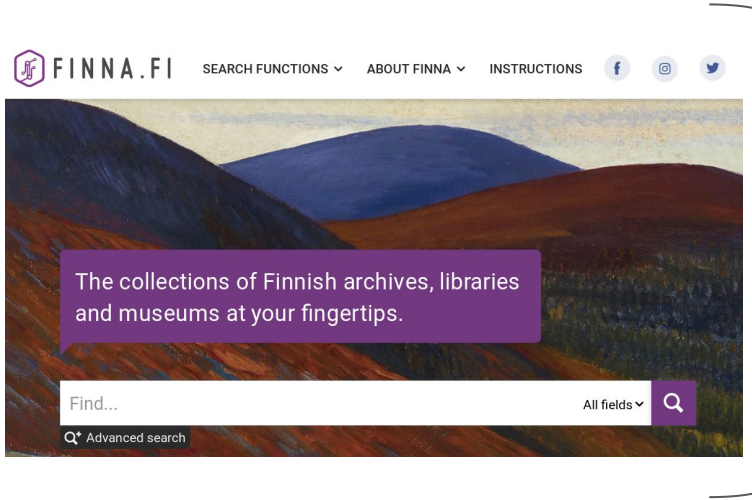
1. Development of Annif





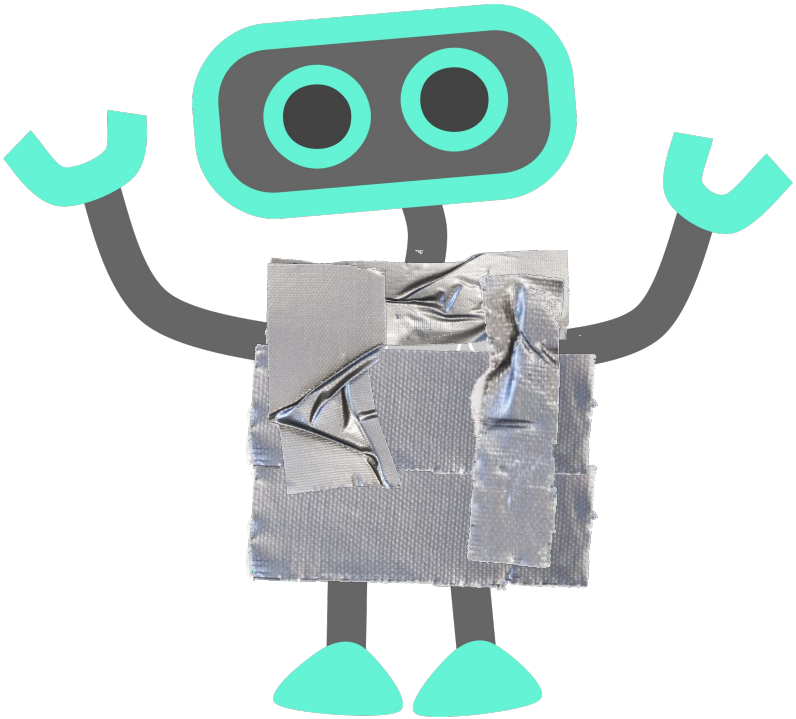
YSO, General Finnish Ontology
with 30,000+ subjects

Machine learning using existing metadata



annif

Early prototype (2017) got people excited

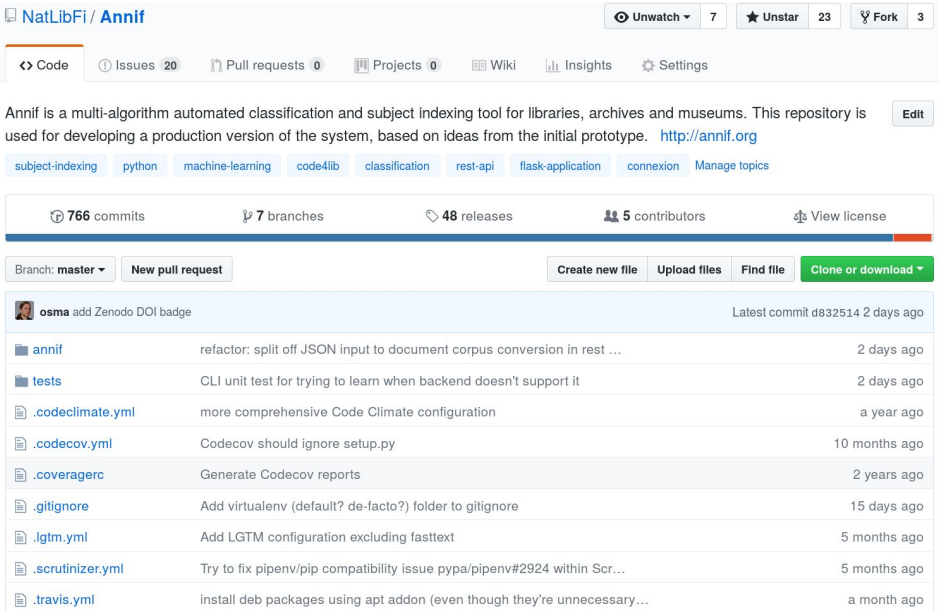


Starting points for Annif implementation (2018 →)

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

annif

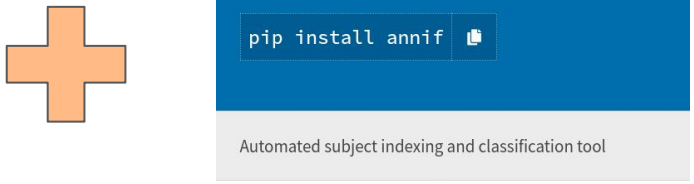
Annif on GitHub



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

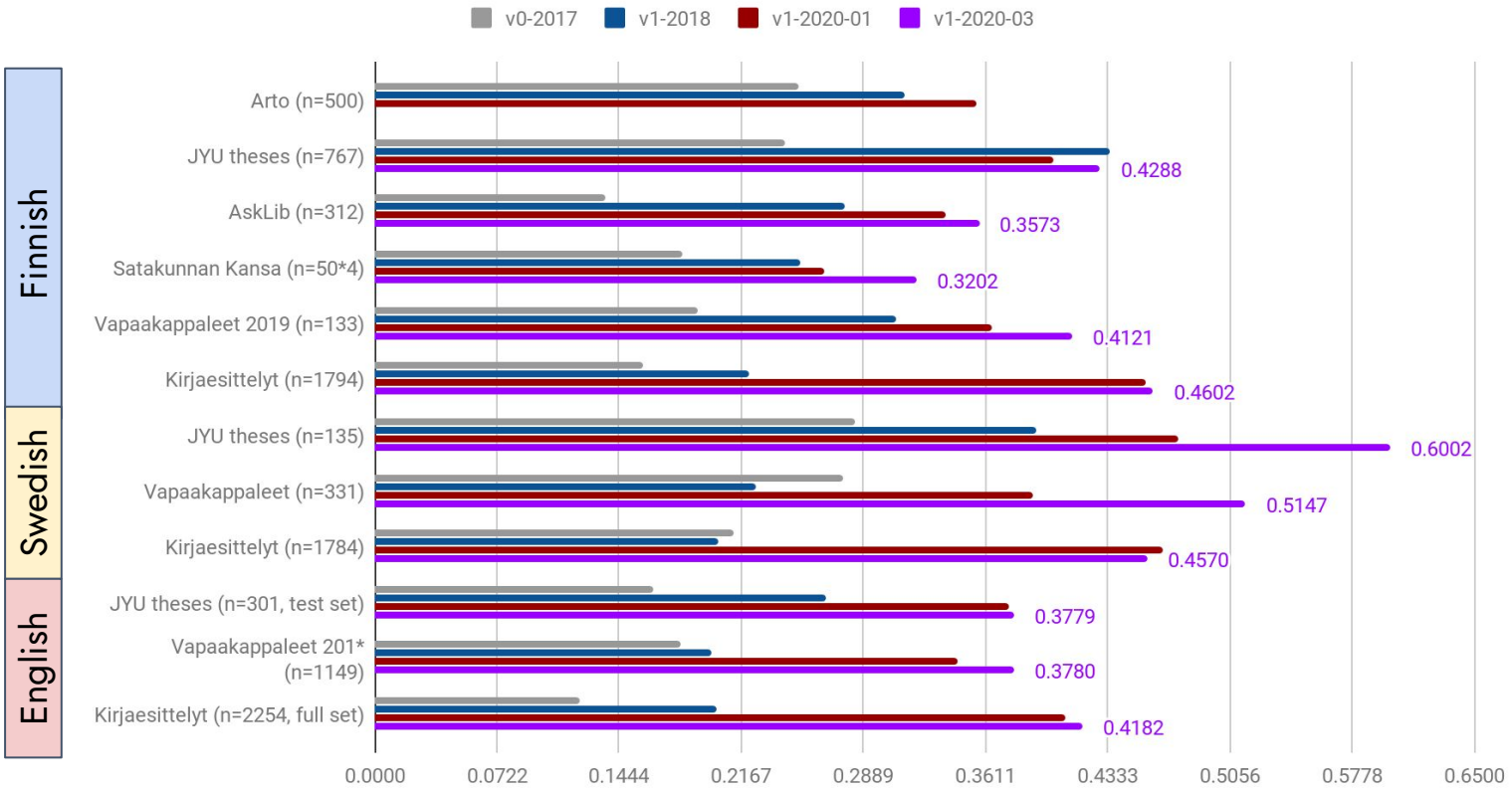
1. Metadata records from Finna.fi discovery system
2. Ask a Librarian question-answer pairs
3. Master's and Doctoral theses from University of Jyväskylä
4. Book descriptions from publishers (via Kirjavälitys Oy)
5. E-books from our electronic deposit system
6. …

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).

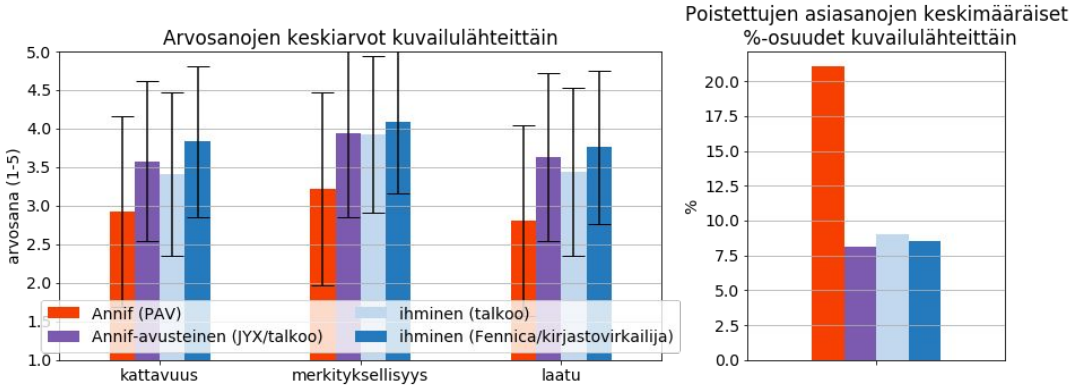


Photo: Mikko Lappalainen.

Lehtinen M., Inkinen J. & Suominen O. (2019). Aaveita koneessa: Automaattisen sisällönkuvailun arviointia Kirjastoverkkopäivillä 2019. *Tietolinja, 2019(2)*. http://urn.fi/URN:NBN:fi-fe2019120445612

Annif-Leiki Comparison at Finnish Broadcasting Company Yle

- Annif vs Leiki (commercial service) tagging compared by 28 human evaluators at Yle

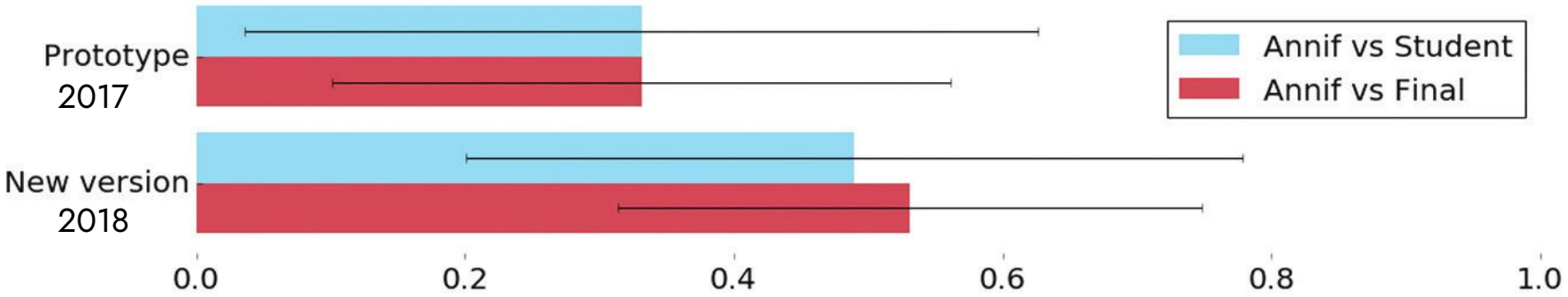
- About 100 Finnish and Swedish articles and their tags
  - business, science, culture, sport

Finnish: Annif **slightly better** than Leiki
Swedish: Annif **substantially** better than Leiki

Suominen, O. & Virtanen, P. Yle meets Annif – an open source tool for automated subject indexing. Presentation at EBU MDN Workshop 2020, 10 June 2020.

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

annif

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

☆ Annif Users

1–30/77

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 2	31. elok.	☆
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/Ann	30. heinäk.	☆

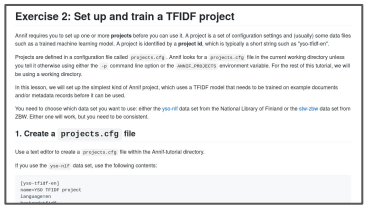
Hands-on Annif tutorial

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



Videos and exercises freely available on YouTube & GitHub!

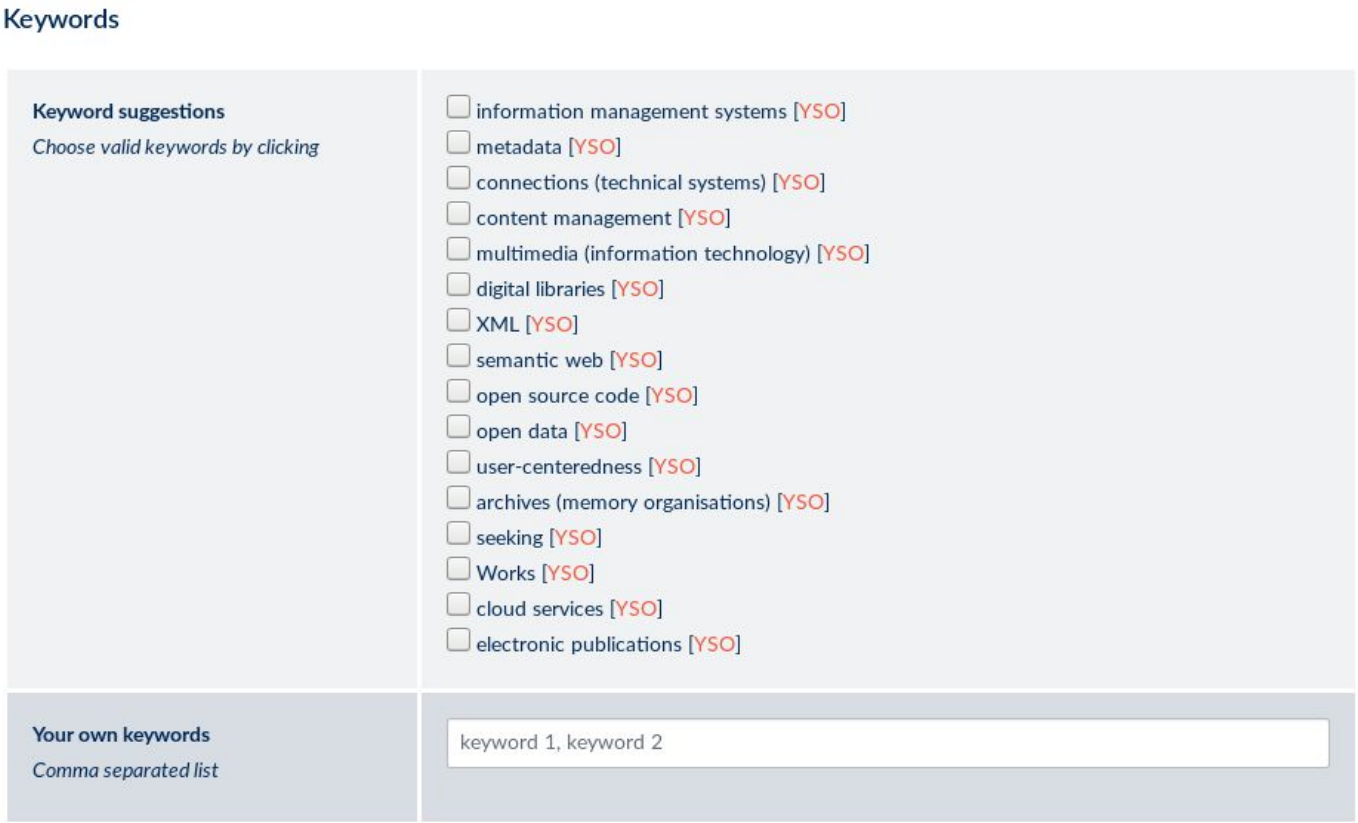
16 40
THE NATIONAL LIBRARY OF FINLAND

ZBW
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*



Implemented using DSpace & GLAMpipe by Ari Häyrinen

*from YSO = General Finnish Ontology

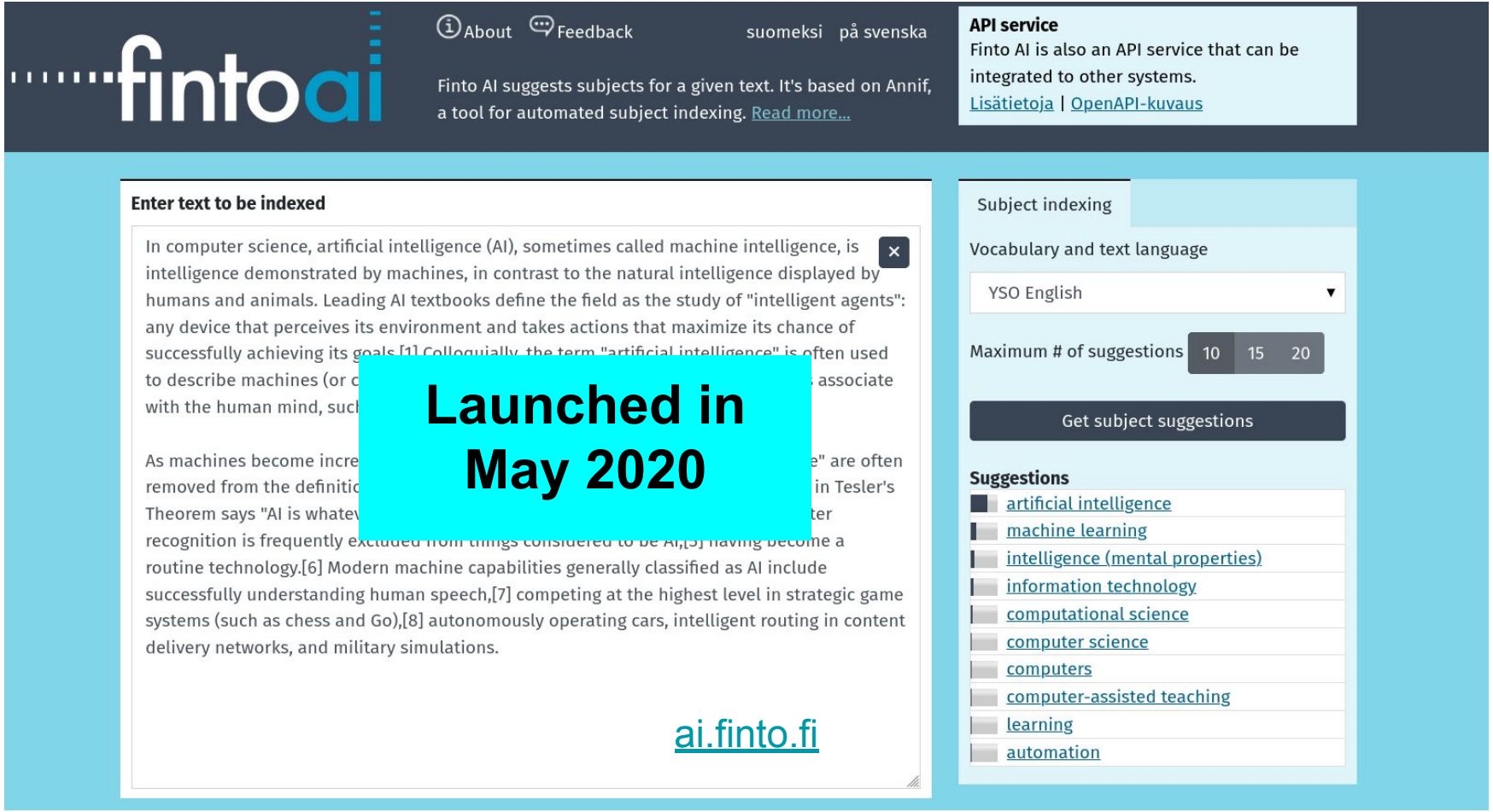
Osuva repository, University of Vaasa
Trepo repository, University of Tampere
Theseus repository, Finnish universities of applied sciences

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.

Finto AI - automated subject indexing tool and API service

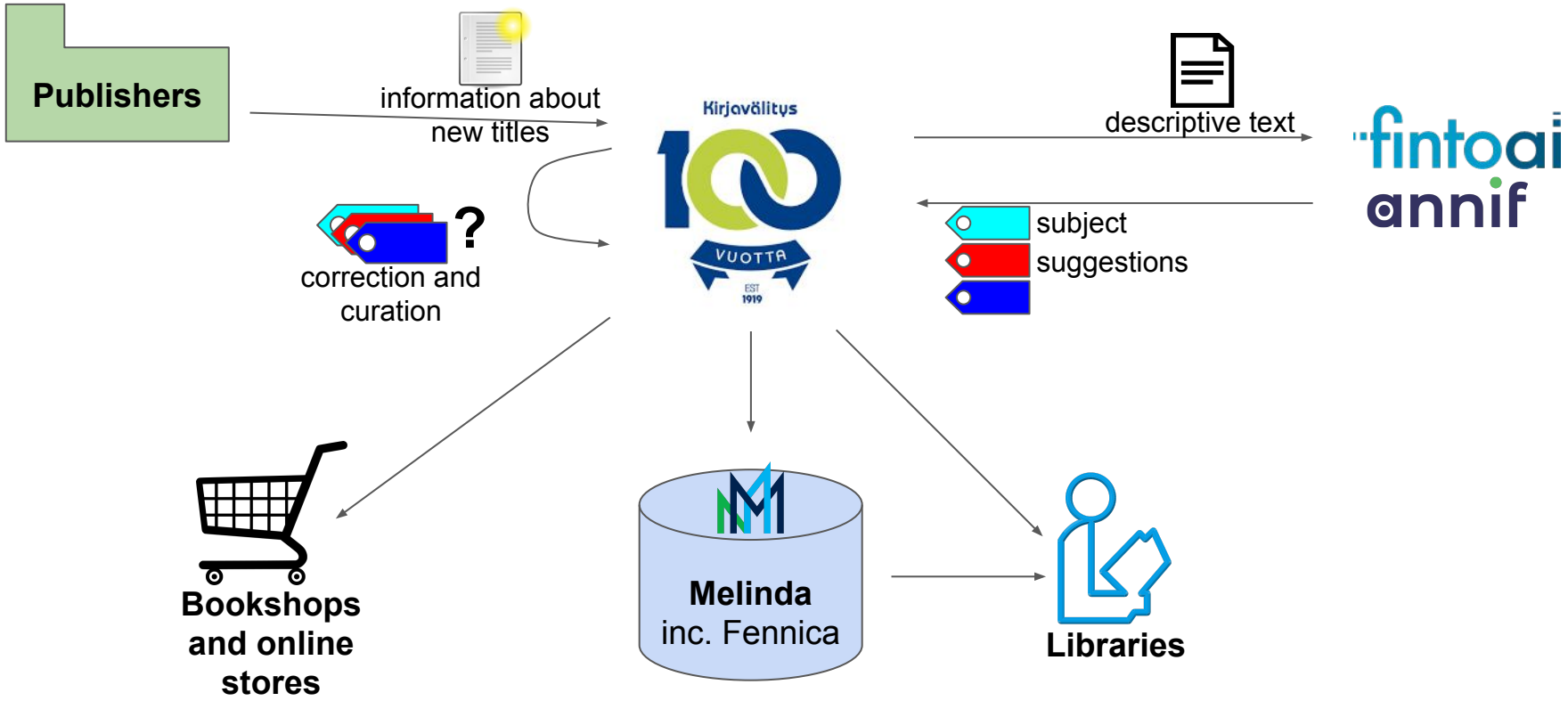


Subject indexing for electronic deposits

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.

Implementation: Erik Lindgren, Mikko Merioksa, Satu Niininen

Kirjavälitys Oy - logistics company serving bookstores and libraries

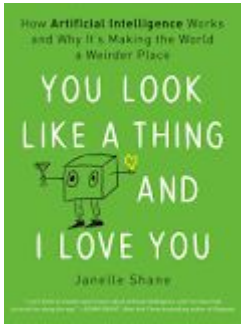


5. Lessons learned

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

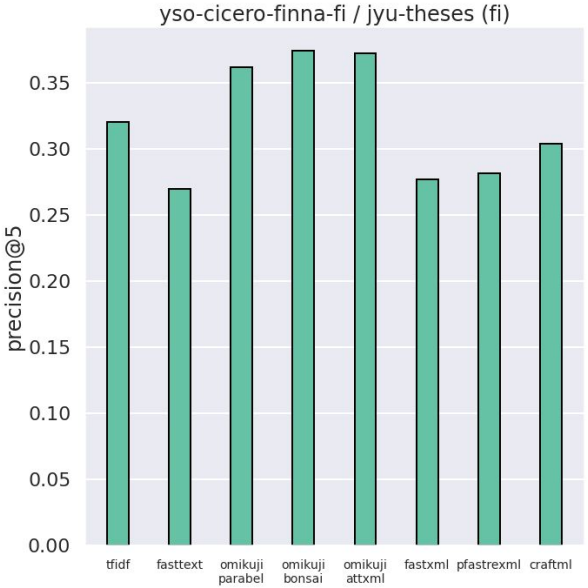
Keyword suggestions
Choose valid keywords by clicking

- 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

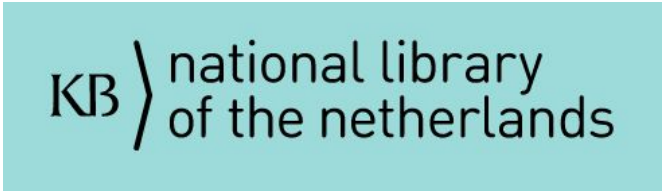
Collaboration is valuable! (1)



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)



[1,2]

[3]

[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-ohjelmatietokanta Ritvassa. Informaatiotutkimus 39 (1), 2020. DOI: https://doi.org/10.23978/inf.88107

Thank you!

Juho Inkinen

Mona Lehtinen

Osma Suominen

annif.org

These slides: **https://tinyurl.com/annif-swib20**