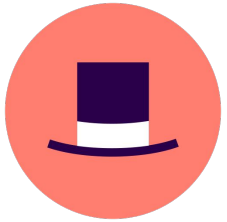
annif tutorial



Introduction to Annif and automated subject indexing

Osma Suominen

DCMI Virtual, 21 September 2020

About me

Twitter:
@OsmaSuominen

LinkedIn:
osmasuominen

GitHub:
@osma

Osma Suominen
Information Systems Specialist, National Library of Finland

finto

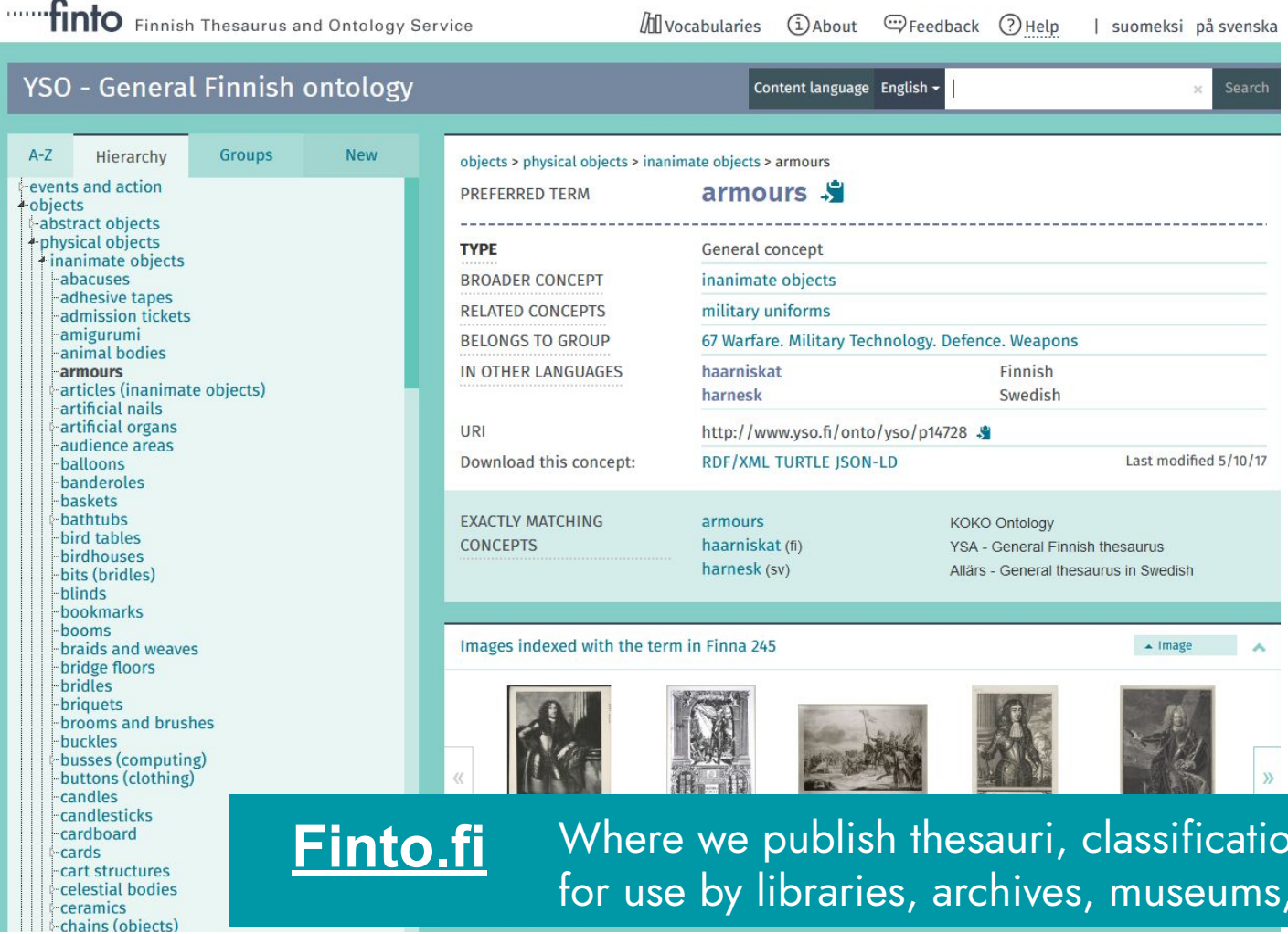
Doctoral thesis "*Methods for Building Semantic Portals*"
Semantic Computing Research Group, Aalto University, 2013
Supervisor Professor Eero Hyvönen

Joined the National Library in 2013
to set up the Finto.fi thesaurus and ontology service

Working on opening up bibiliographic metadata as Linked Data (Fennica-LD) and automated subject indexing (Annif, Finto AI)

Open source software projects e.g.:

- Skosify - Validation and QA tool for SKOS vocabularies
- Skosmos - SKOS vocabulary publishing tool
- Annif - Tool for automated subject indexing and classification



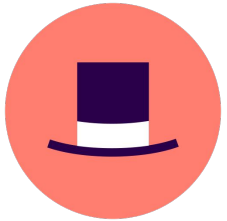
Subject indexing vocabularies:

General Finnish Ontology YSO
(trilingual fi, sv, en, with 30,000+ concepts)

KOKO Ontology

...and many more

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



Subject indexing

a.k.a. topic indexing, topic assignment, term assignment

~ tagging

~ multi-label classification

Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

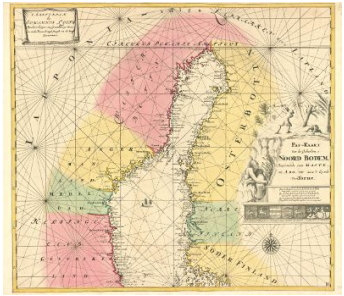
Finlands oavhängighet.

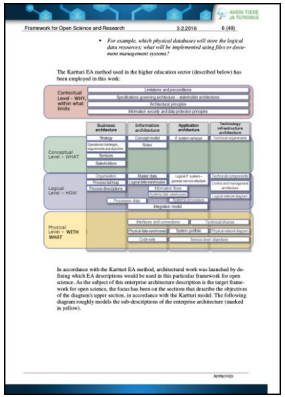
Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.

Våldsdåd av arbetare.

Återupptagandet av trafiken Åbo—Stockholm.

VENNY SOLDAN-BROFELDT.





Framework for Open Science and Research 3.2.2016 6 (49)

- *For example, which physical databases will store the logical data resources; what will be implemented using files or document management systems?*

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

In accordance with the Kartturi 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 Kartturi model. The following diagram roughly models the sub-descriptions of the enterprise architecture (marked in yellow).

Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

Finlands oavhängighet.

Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.

Vår speciella korrespondent [illegible] i Helsingfors meddelar:

Den förda linjen vid plenar- [illegible] av. I samtliga att ge de hemvändande programmet beredskap. Vanliga till att [illegible] och [illegible] ordning av [illegible] upprätthållas.

Förslagen lyda:

Med erkännande därav att regeringen till lösandet av den uppkomna [illegible] den med förslag till ny regeringsform, riksdag att [illegible] att [illegible] och [illegible] [illegible] [illegible]

Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

Finlands oavhängighet.

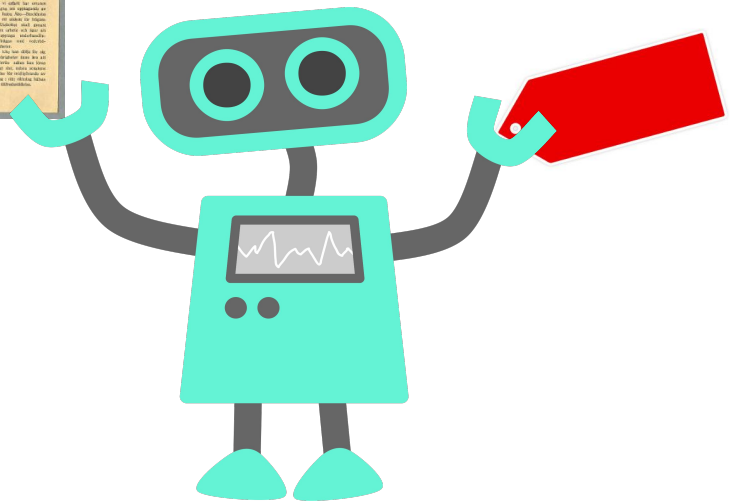
Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.

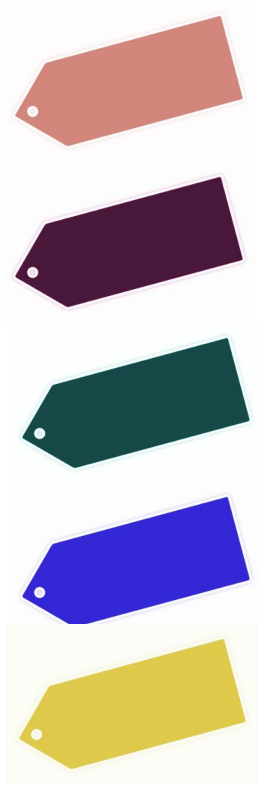
Våldsdåd av arbetare.

Återupptagandet av trafiken Åbo—Stockholm.

Senaten tillsätter ett utskott.

Pris 25 penni.

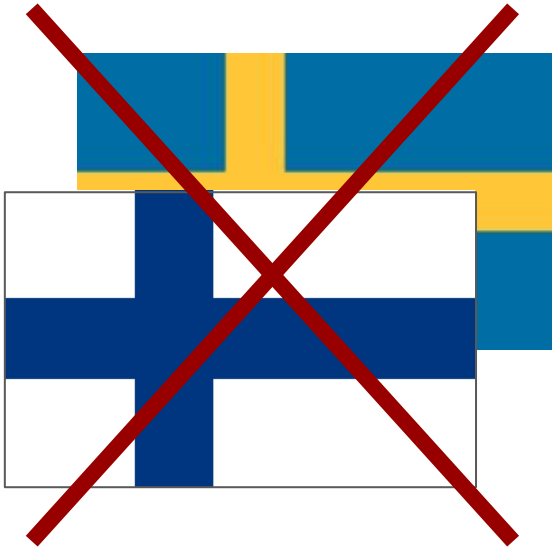




YSO, General Finnish Ontology
with 30,000+ subjects

OPEN CALAIS

THOMSON REUTERS



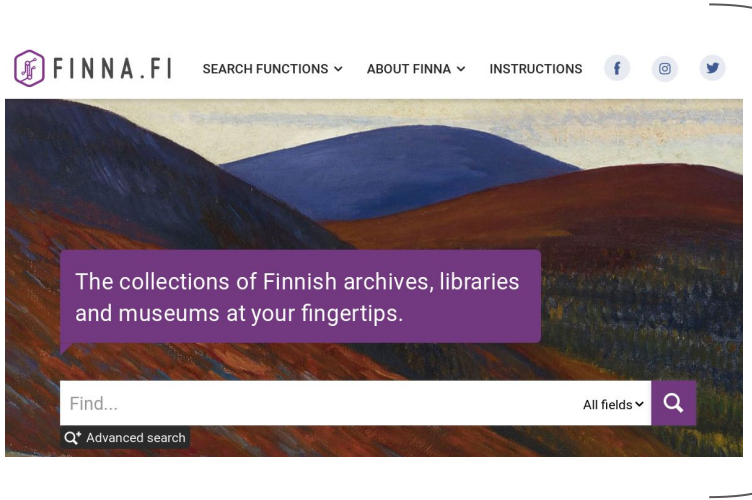
YSA
Allärs

YSO
KOKO

€£$

black box

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"

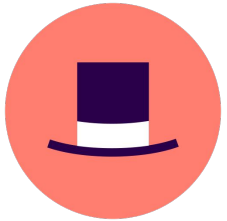
Lexical approaches need comparatively little training data.

associative approaches (e.g.: TF-IDF, fastText, Omikuji)

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

Associative approaches need a lot more
training data in order to cover each subject.

Algorithms used in Annif

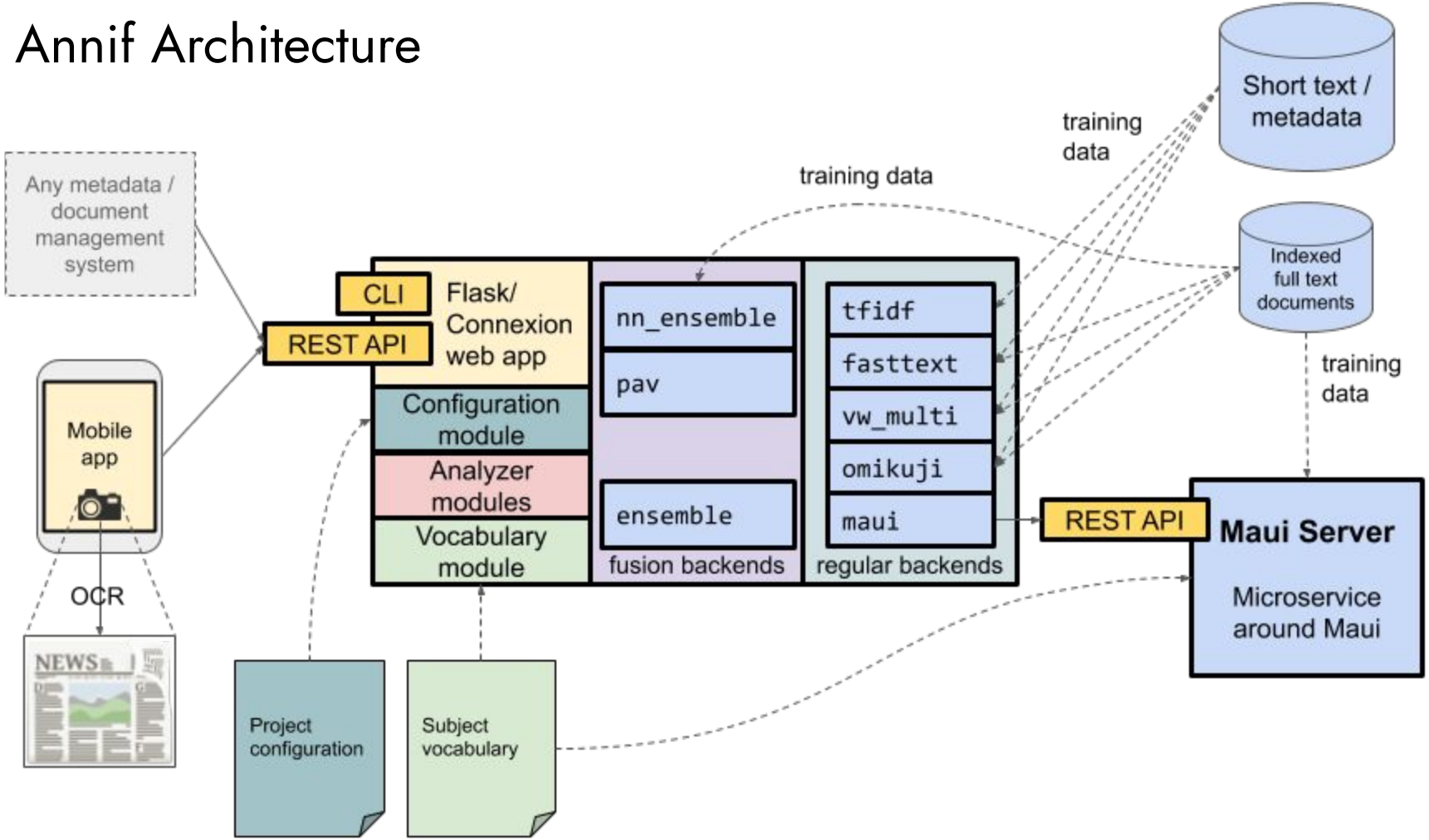


lexical	Maui (using the Maui Server REST API) Maui is a lexical tool for automated indexing
associative	TF-IDF similarity (implemented with the Gensim Python library) baseline bag-of-words similarity measure and vector space model
	fastText (by Facebook Research) uses word embeddings and simulates a deep neural network architecture
	Parabel and Bonsai (implemented with the Omikuji Python library) tree-based algorithms for extreme multi-label classification (i.e., when the set of subjects is huge)

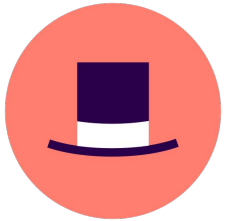
Implemented as Annif backends – see the Annif wiki documentation for details about each backend

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

Annif Architecture



Accessing Annif



Command line interface

- setup and administration
- training models
- testing and evaluating models
- bulk indexing of documents

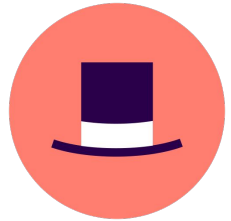
Web user interface

- interactive testing of models

REST API

- integrating Annif services to other systems

API access example



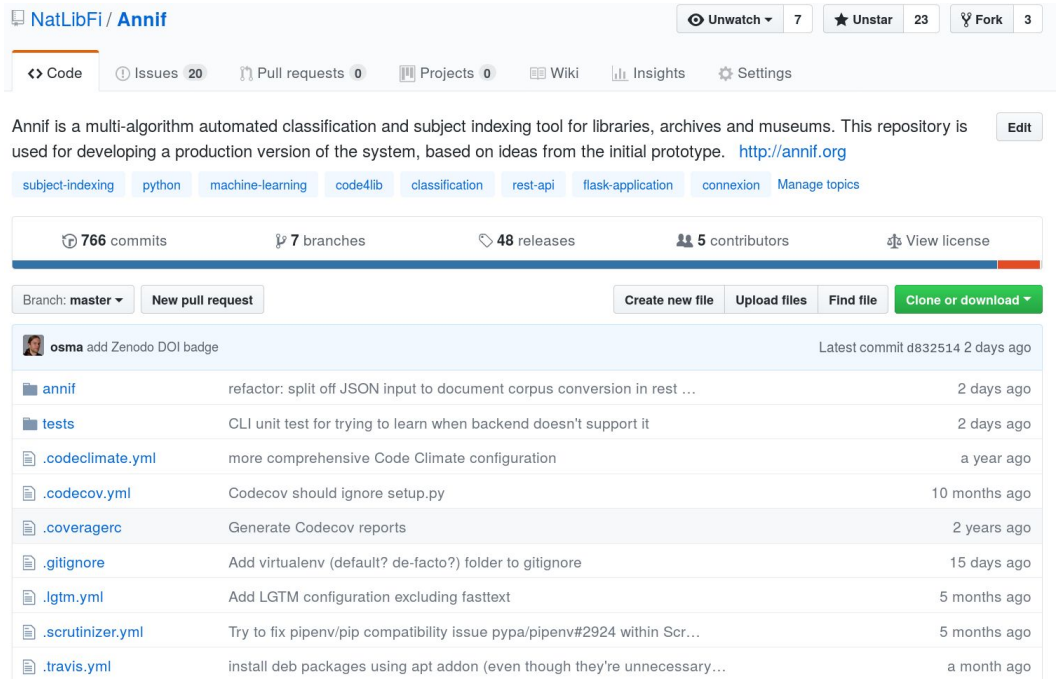
“The quick brown fox jumped over the lazy dog.”

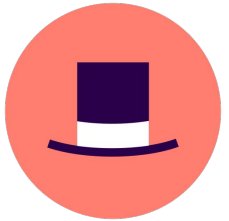
suggest

annif

api.annif.org

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**”}
]





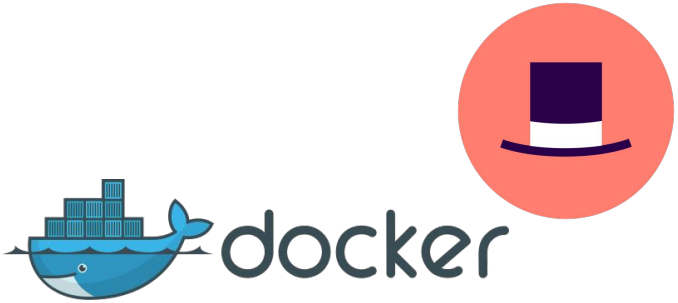
Annif on GitHub

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/



quay.io/natlibfi/annif

annif 0.47.1

```
pip install annif
```

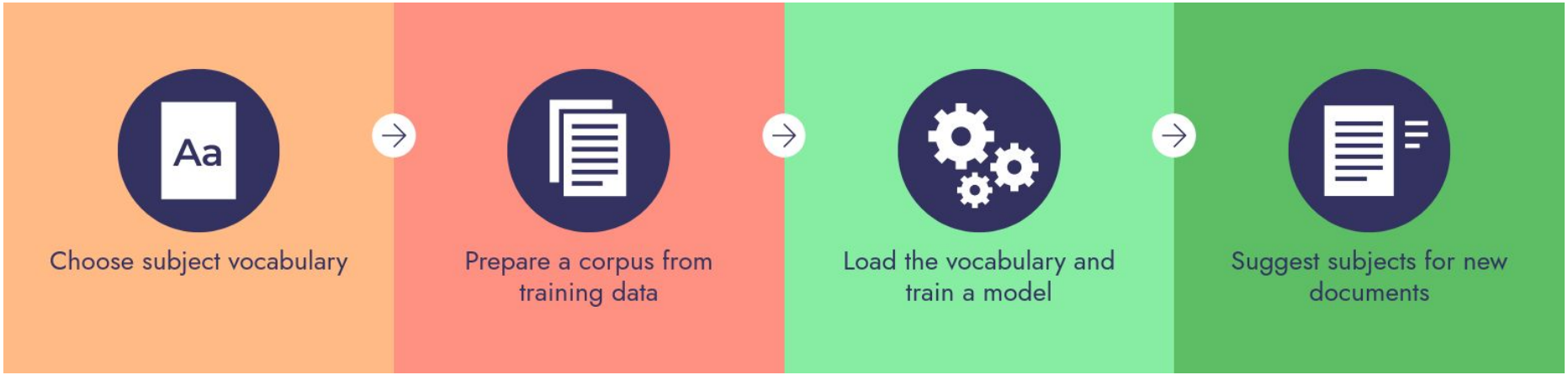
Automated subject indexing and classification tool

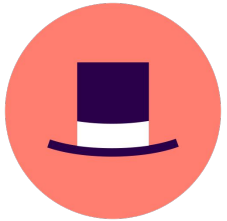
```
jmminkin@lx8-9811-008: /home/local/jmminkin/git/Annif
jmminkin@lx8-9811-008: /home/local/jmminkin/git/Annif 99x35
(Annif) jmminkin@lx8-9811-008:/home/local/jmminkin/git/Annif$ annif
Usage: annif [OPTIONS] COMMAND [ARGS]...

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!





Demonstration of Annif

1. Load STW Thesaurus for Economics
2. Train a small model on metadata from the EconBiz portal
3. Test the model using the Web UI

Form for testing at annif.org

TRY THE DEMO!

Manually indexing documents for subject-based access is a very labour-intensive intellectual process. machine could perform similar subject indexing much faster. In this series of presentations and demonstrations, we will show practical examples of automated subject indexing and discuss how such systems can be evaluated.

In the first part of this presentation, Osma Suominen will introduce the general idea of automated subject indexing using a controlled vocabulary such as a thesaurus or a classification system; and the open source automated subject indexing tool Annif, which integrates several different machine learning algorithms for text classification. By combining multiple approaches, Annif can be adapted to different settings. The tool can be used with any vocabulary; and, with suitable training data, documents in many different languages may be analysed. Annif is both a command line tool and a microservice-style API service which can be integrated with other systems. We will demonstrate how to use Annif to train a model using metadata from an existing bibliographic database and how it can then provide subject suggestions for new, unseen documents.

In the second part of the presentation, Koraljka Golub will discuss the topic of evaluating automated subject indexing systems. There are many challenges in evaluation, for example the lack of gold standards to compare against, the inherently subjective nature of subject indexing, relatively low inter-indexer consistency in typical settings, and dominating out-of-context, laboratory-like evaluation approaches.

PROJECT (VOCABULARY AND LANGUAGE)

YSO NN Ensemble English

MAX # OF SUGGESTIONS

10 15 20

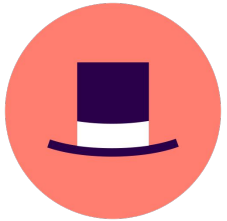
Get suggestions

annif

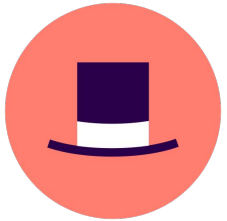
SUGGESTED SUBJECTS

- indexing
- information retrieval
- subject cataloging
- evaluation
- documentation
- lists of subject headings
- artificial intelligence
- classification systems
- automation
- classification

Questions on the introductory part?



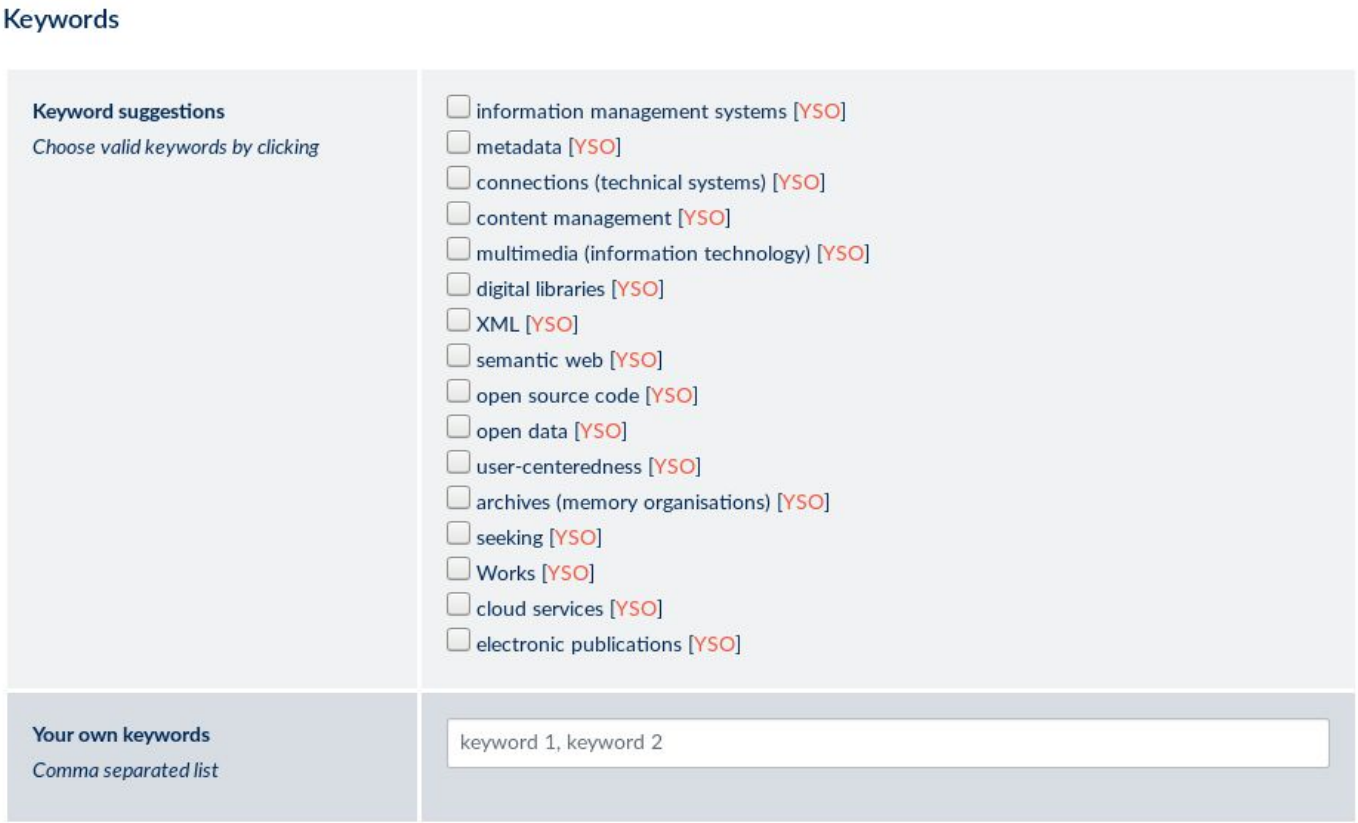
Next up:
Koraljka Golub: Evaluating automated subject indexing
Annemieke Romein & Sara Veldhoen: Case study on applying Annif on legal texts
Osma Suominen: Where Annif is used, evaluation results & hands-on tutorial



Annif used in production

JYX repository, University of Jyväskylä

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



Implemented using DSpace & GLAMpipe by Ari Häyrinen

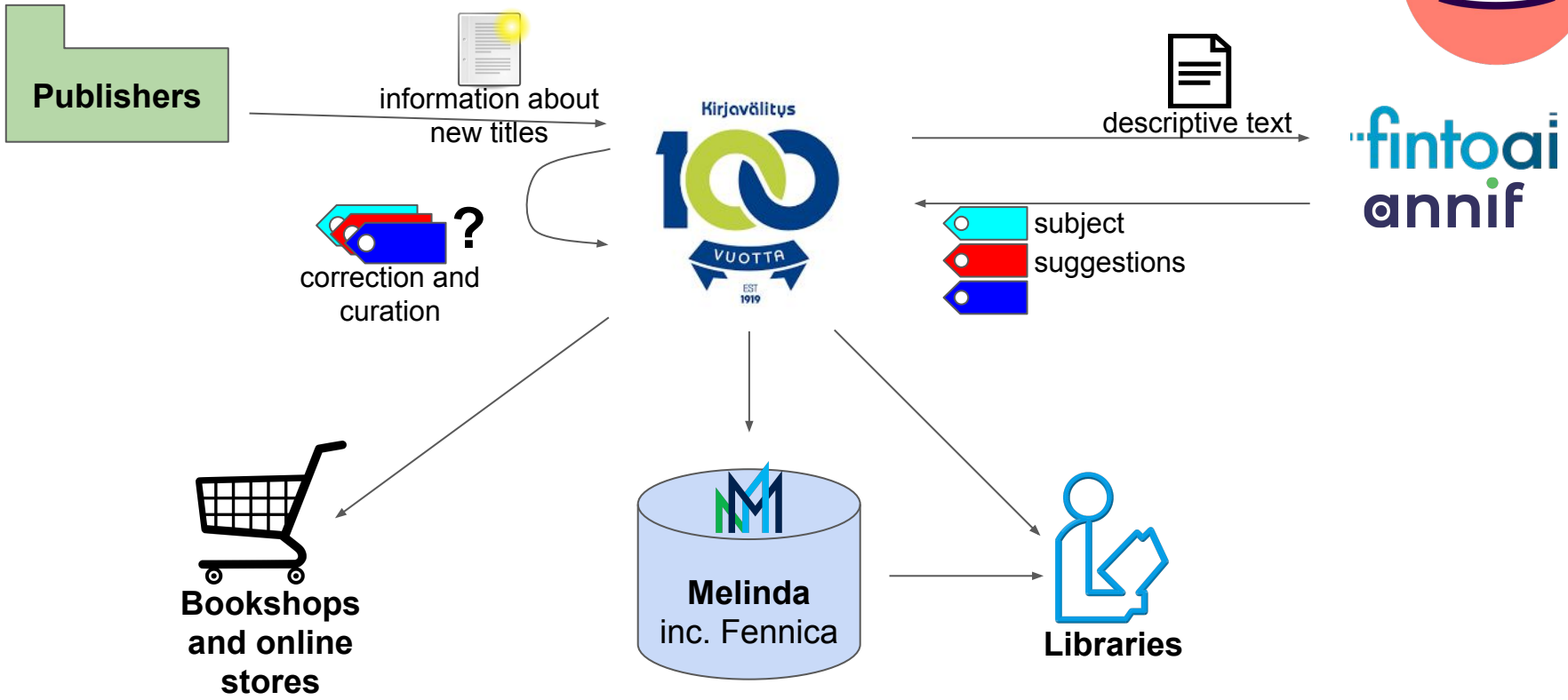
*from YSO = General Finnish Ontology

Osuva repository, University of Vaasa

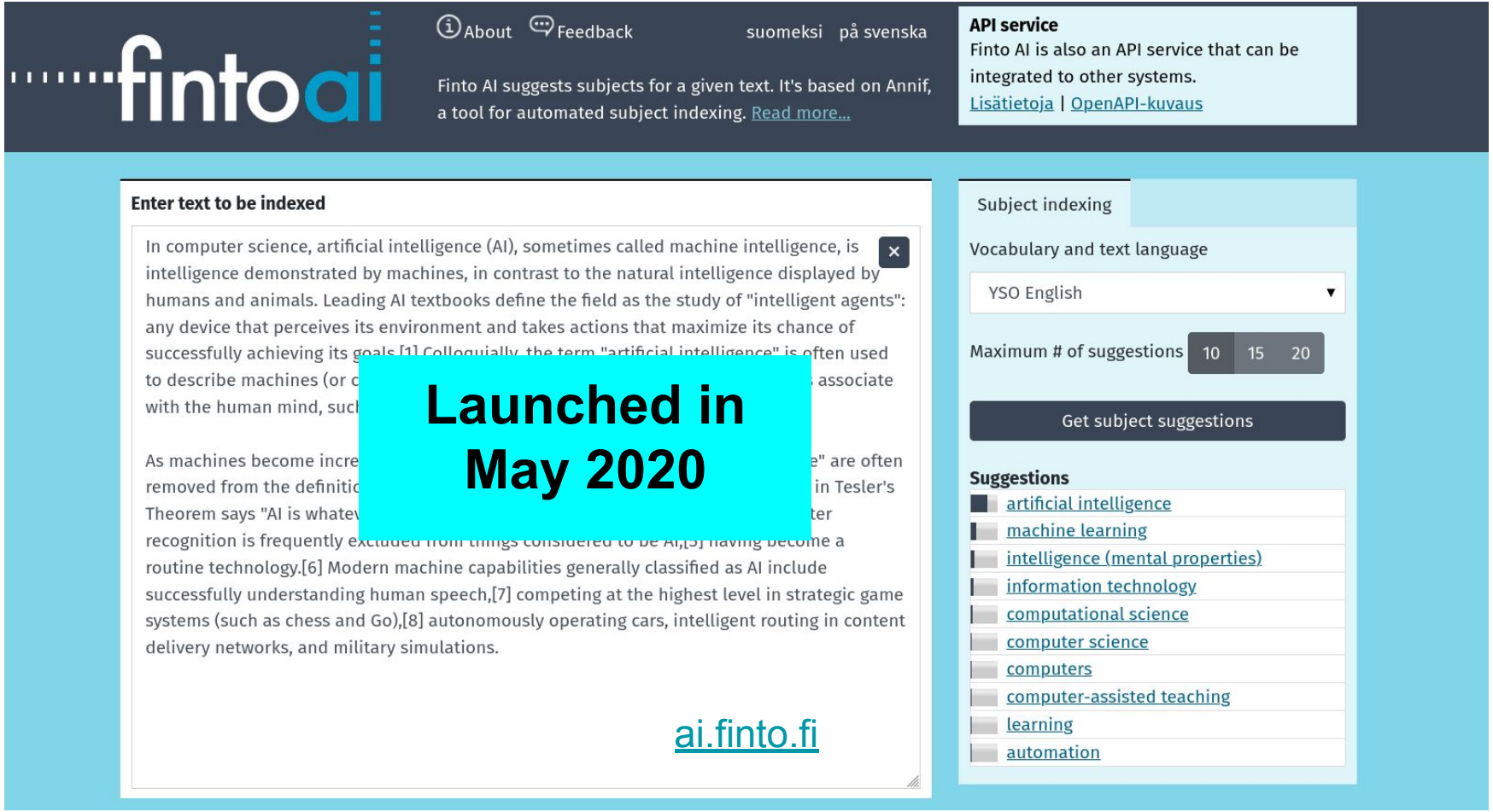
Same idea as JYX: students upload their theses,
Annif suggests subjects

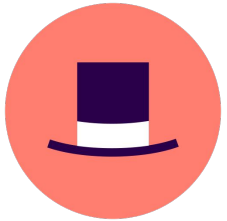
Pilot started
2.3.2020,
implementation by
Anis Moubarik.

Kirjavälitys Oy - logistics company serving bookstores and libraries



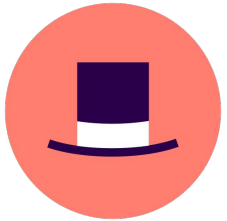
Finto AI - automated subject indexing tool and API service





Evaluating algorithms used in Annif

Evaluation approaches (Golub et al. 2016), **emphasis** mine



1. Evaluating indexing quality directly through **assessment by an evaluator** or by **comparison with a gold standard**.
2. Evaluating indexing quality directly **in the context of an indexing workflow**.
3. Evaluating indexing quality indirectly through retrieval performance.

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.

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).

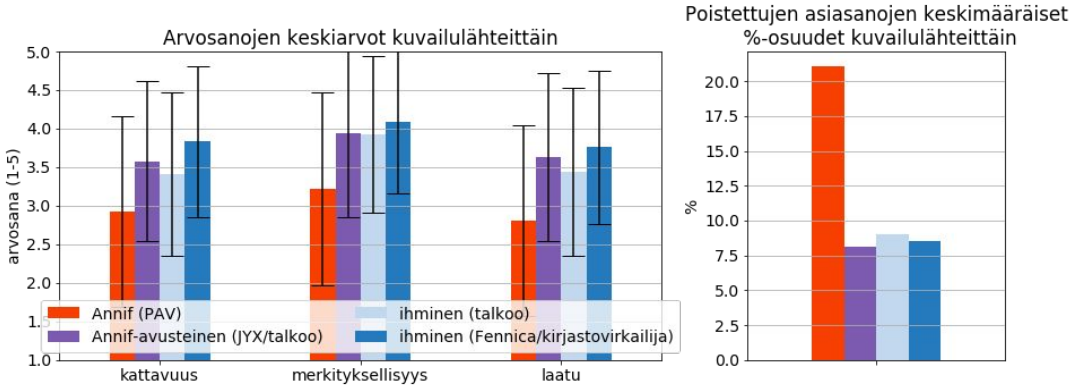


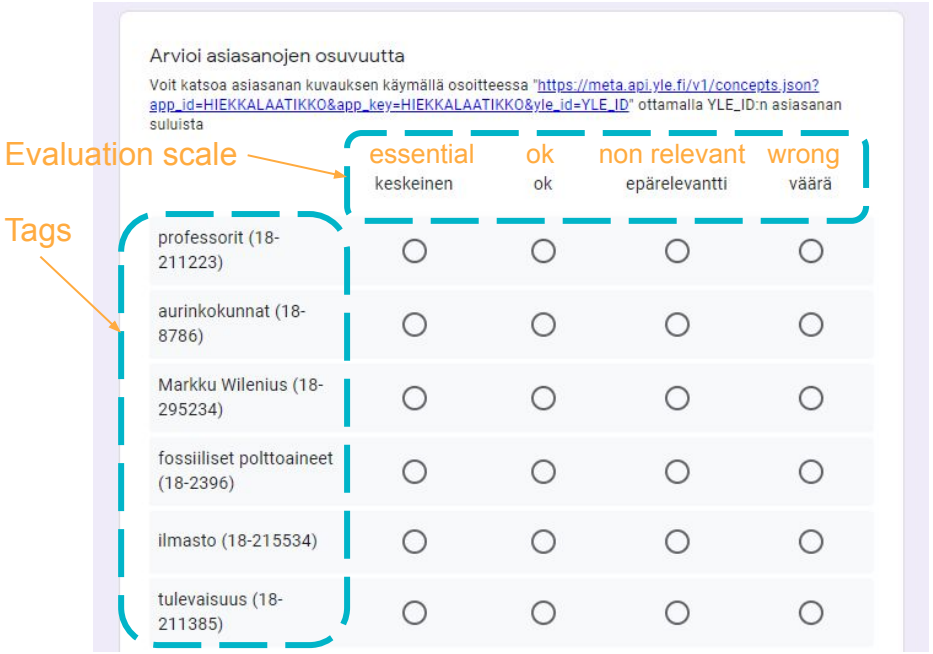
Photo: Mikko Lappalainen.

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



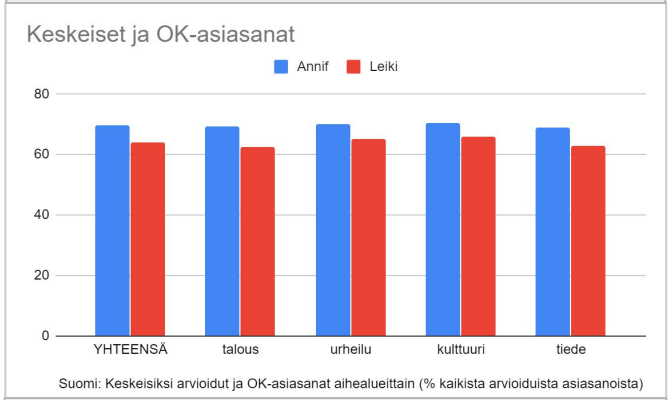
Suominen, O. & Virtanen, P. Yle meets Annif – an open source tool for automated subject indexing. Presentation at EBU MDN Workshop 2020, 10 June 2020.

Comparison: Overall Results / Finnish

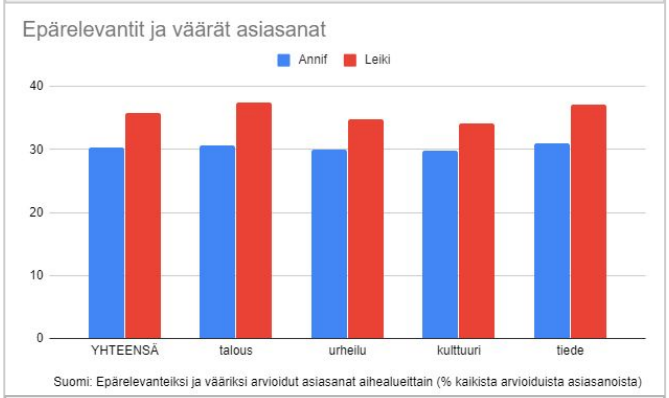
Annif performed slightly better than Leiki
= more essential + ok, less not relevant + wrong tags

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

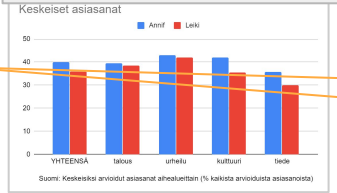
Essential + ok (% of all tags)
TOTAL - business - sport - culture - science



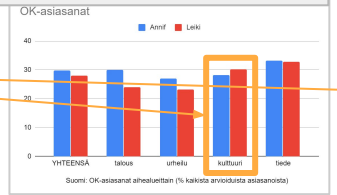
Not relevant + wrong (% of all tags)
TOTAL - business - sport - culture - science



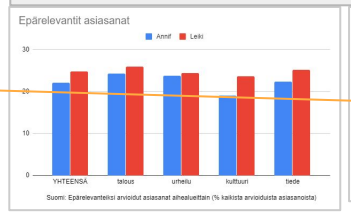
essential



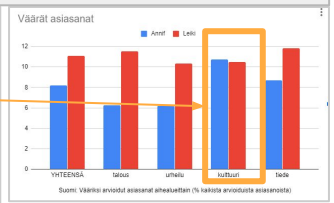
ok



non relevant



wrong



Comparison: Overall Results / Swedish

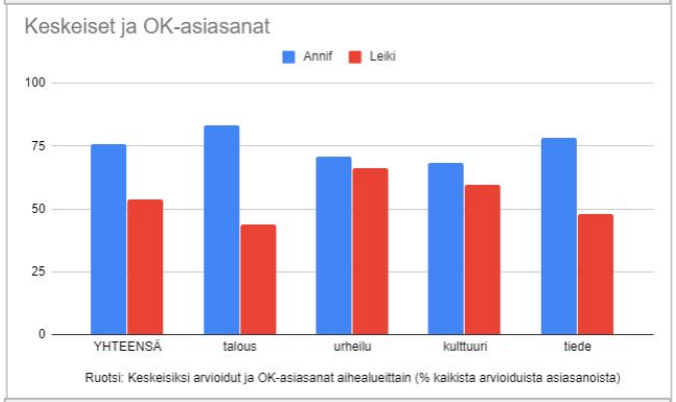
Annif performed better than Leiki in all subject areas
= more essential + ok, less not relevant + wrong tags

Differences bigger than in Finnish

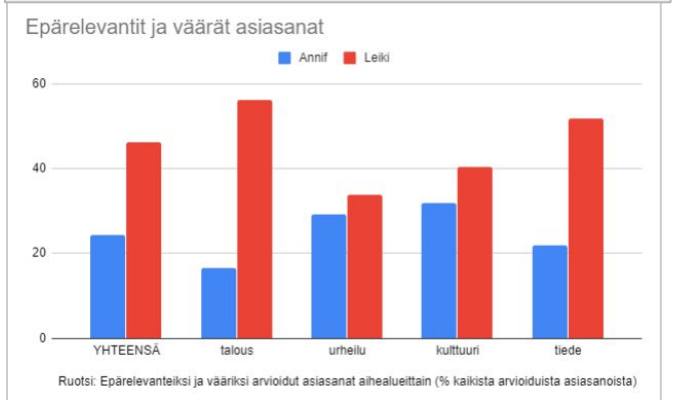
Biggest differences in business and science

Reasons?

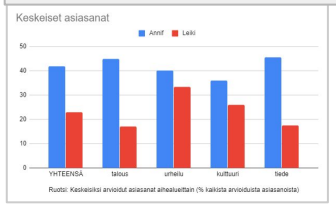
Essential + ok (% of all tags)
TOTAL - business - sport - culture - science



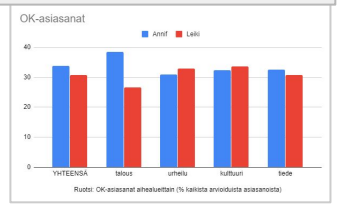
Not relevant + wrong (% of all tags)
TOTAL - business - sport - culture - science



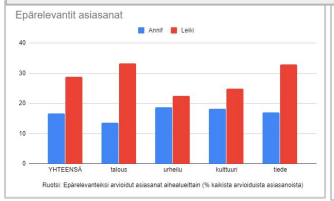
essential



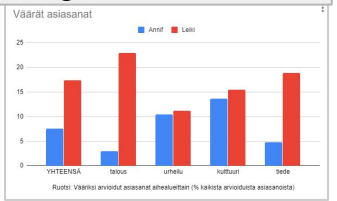
ok



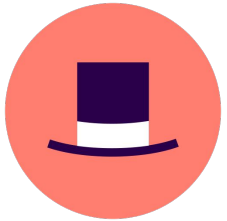
non relevant



wrong



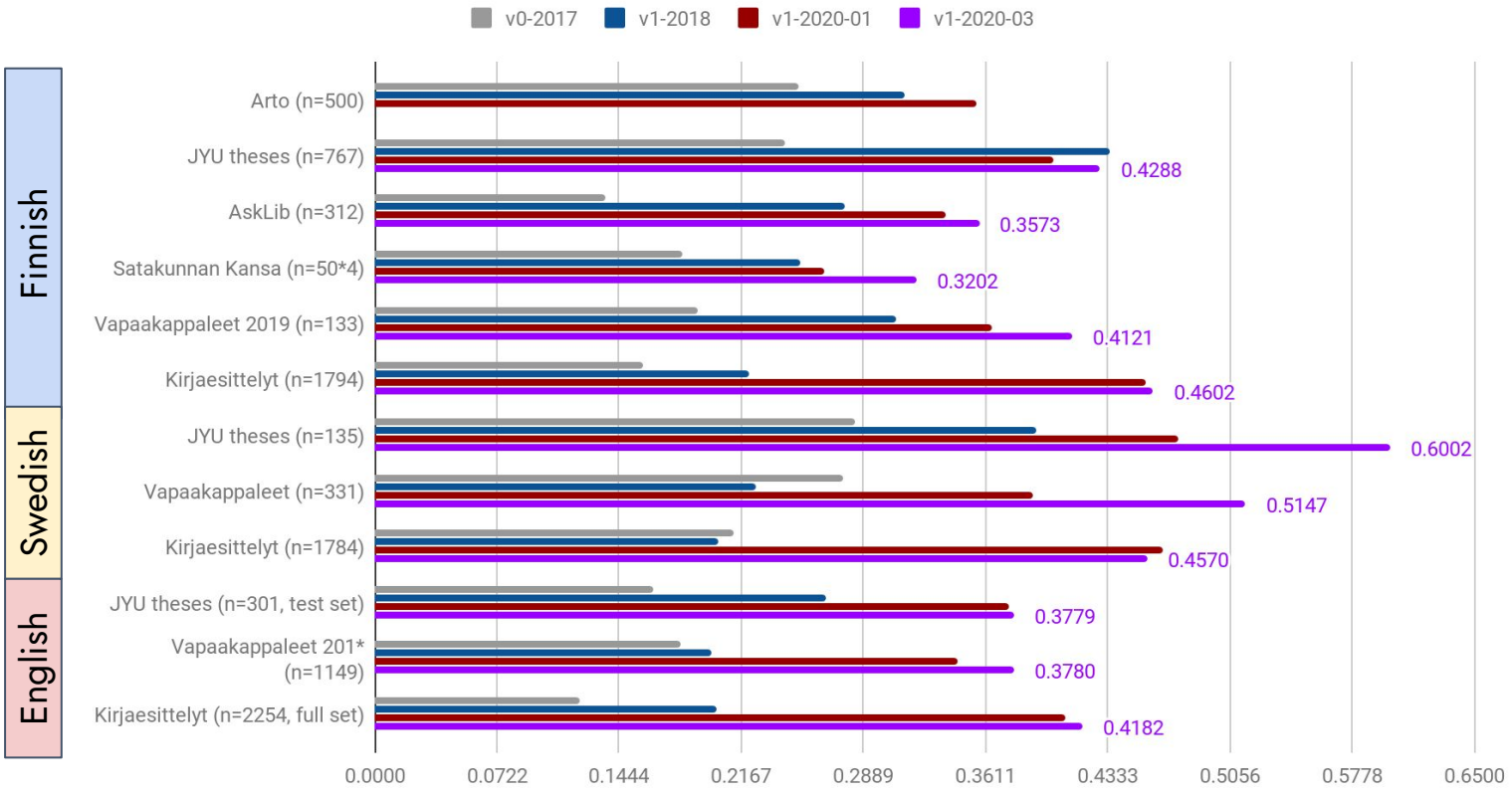
Precision, recall and F1 score



- **Precision:** fraction of the correct subjects among the subjects suggested
  "How many of the suggested subjects are actually correct?"
- **Recall:** fraction of all correct subjects that were actually suggested
  "How many of those subjects that should be suggested have actually been suggested?"
- The **F1 score** is the harmonic mean between precision and recall
  (i.e., a way of combining precision and recall values into one **similarity score**).

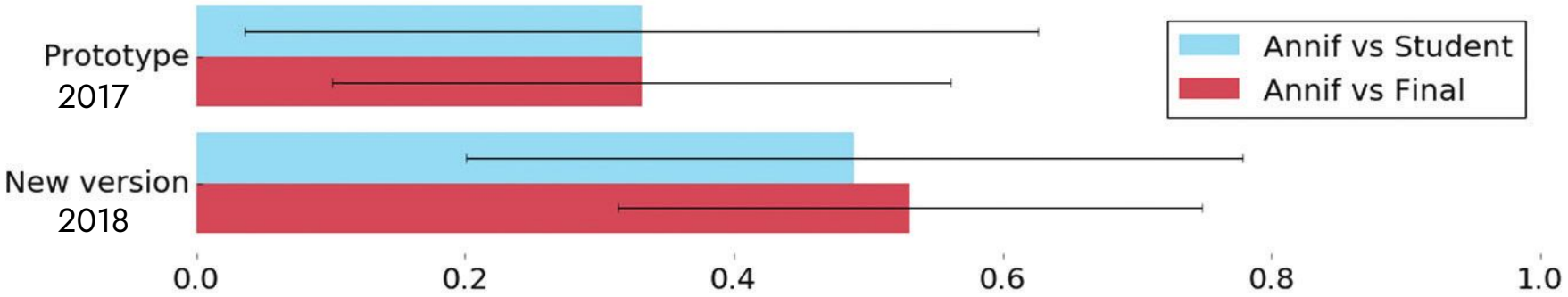
Comparison to “gold standard”

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



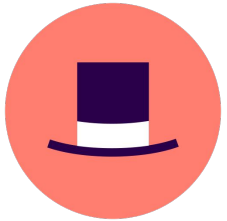
Evaluating in the context of an indexing workflow

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)

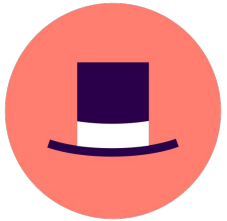


Suominen, O., 2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, 29(1), pp.1–25. DOI: http://doi.org/10.18352/lq.10285

Lessons from evaluation



- The different evaluation approaches are complementary. Not a good idea to look at just a single measure.
- Improved quality of automated subject indexing over time
  - better training and evaluation data
  - better algorithms: Omikuji, neural network ensemble
- Continuous process: it never stops…



Hands-on tutorial

for those who want to learn to use Annif themselves

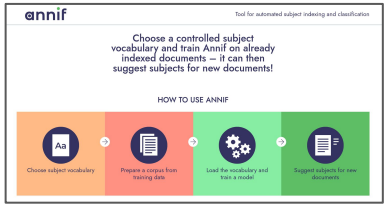
all materials freely available on GitHub & YouTube

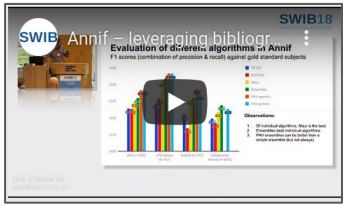
1

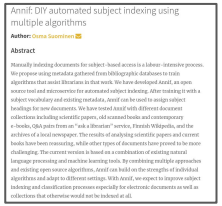
you are here

Understand what Annif is

Study the website annif.org,
watch a presentation about it,
or read the LIBER Quarterly paper.



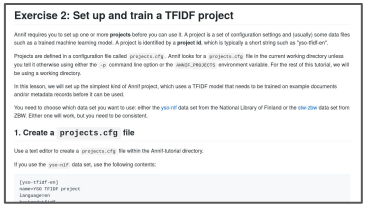




2

Complete this hands-on tutorial

Watch the videos, install Annif, and
complete the exercises as far as you can,
on your own time.

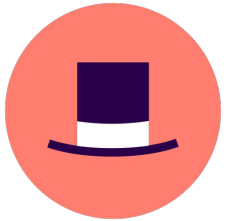


3

Join an online session (optional)

In the online sessions, you can ask questions,
get help and discuss what you've learned.
Registration required.

Videos




annif tutorial
Introduction to the online hands-on tutorial


annif tutorial
Install Annif: usage in VirtualBox


annif tutorial
Install Annif: Docker image


annif tutorial
Install Annif: Linux native install


annif tutorial
Data sets for this tutorial


annif tutorial
TFIDF project


annif tutorial
A little bit about algorithms


annif tutorial
Web UI


annif tutorial
Metrics & evaluation


annif tutorial
Install Maui Server:

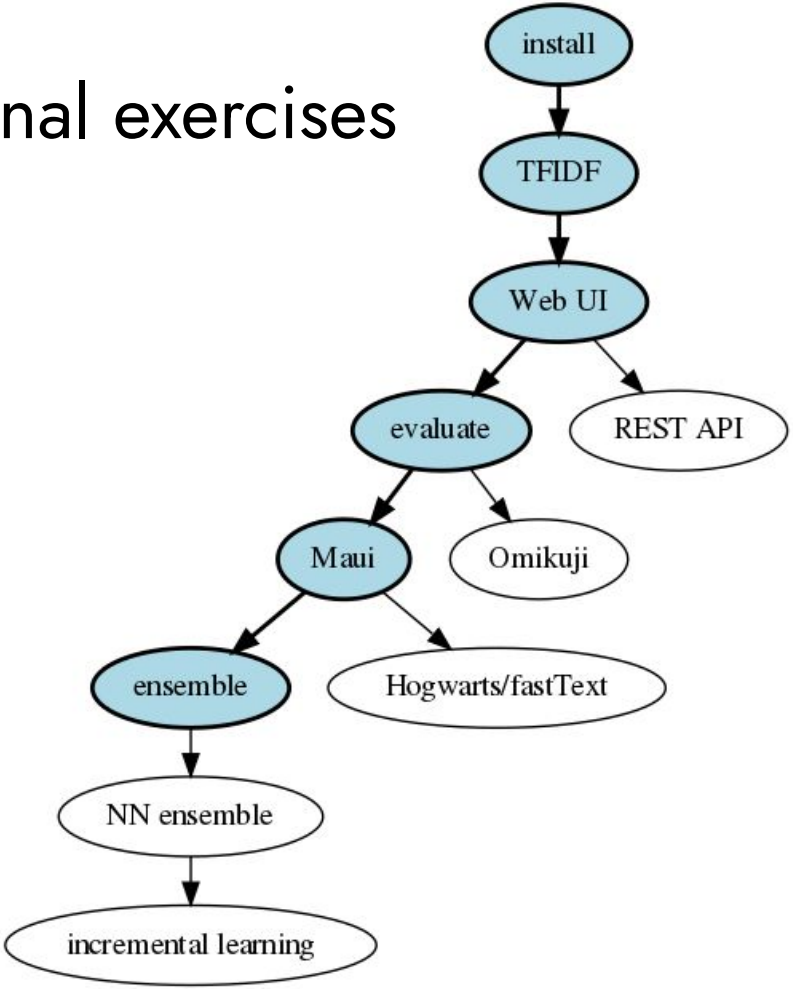

annif tutorial
Install Maui Server:

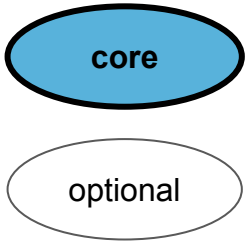

annif tutorial


annif tutorial
Ensembles


annif tutorial

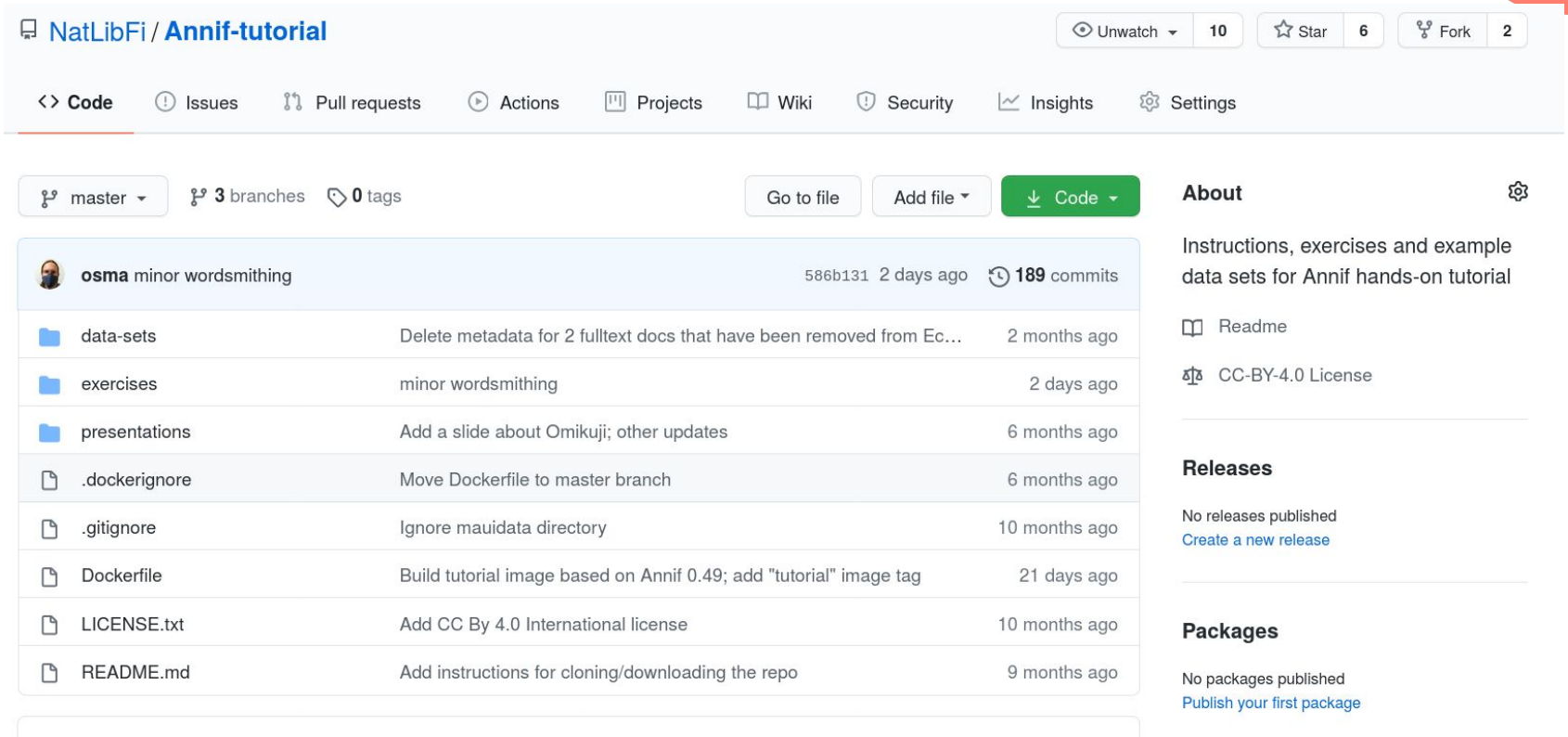
Core and optional exercises



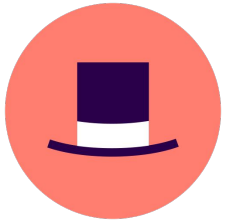


Annif-tutorial GitHub repository

the main resource for the hands-on tutorial



Online help sessions



Only register if you have watched the videos and tried to complete the exercises!

- **Friday 9 October 2020**, 07:00-09:30 UTC
  Registration starts on Friday 25 September

- **Wednesday 21 October 2020**, 15:00-17:30 UTC
  Registration starts on Wednesday 7 October

For more information, see the Annif-tutorial GitHub repository

Automated Subject Indexing IG

Name: Automated Subject Indexing IG

Type:

Status: active

Charter: This group will focus on automated and semi-automated solutions for subject-based information organization of digital collections, i.e., the extraction of semantic features from textual data with methods from areas such as natural language processing and machine learning, and ways of integrating those solutions into productive subject indexing systems.

Moderator/Chair: Koraljka Golub
Anna Kasprzik
Osma Suominen

Established: 2019-04-04

DCMI Automated Subject Indexing IG online meetup
Friday November 6 2020, 09:00-10:30 UTC

Join the mailing list for details

The activities of the group include:

- exchanging ideas on possible research approaches, workflows, algorithms and best practices for automated subject indexing
- collecting and providing information on existing tools, corpora and other resources that are openly available
- working on reusable and interoperable open source tools and coordinating the various development efforts of do-it-yourself solutions within the community
- issuing recommendations for the standardization of APIs and file formats for data exchange, and for more detailed metadata schemata in order to support and document automated methods in subject indexing.

Anyone, not just DCMI members, can join the group. Please subscribe to the autosubject-ig mailing list if you want to take part in activities of the group.

https://www.dublincore.org/groups/automated_subject_indexing_ig/

Thank you!

Juho Inkinen

Mona Lehtinen

Osma Suominen

annif.org

These slides: **https://tinyurl.com/annif-dcmi2020**