Annif

leveraging bibliographic metadata for automated subject indexing and classification

Osma Suominen

SWIB18, Bonn, 28 November 2018

Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

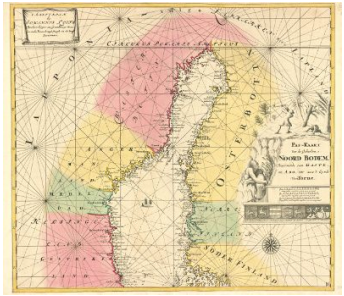
Finlands oavhängighet.

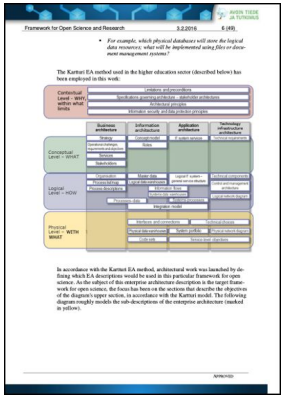
Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.

Våldsdåd av arbetare.

Återupptagandet av trafiken Åbo—Stockholm.

VENNY SOLDAN-BROFELDT.





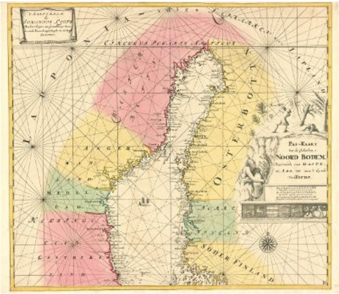
Framework for Open Science and Research 3.2.2016 6 (49)

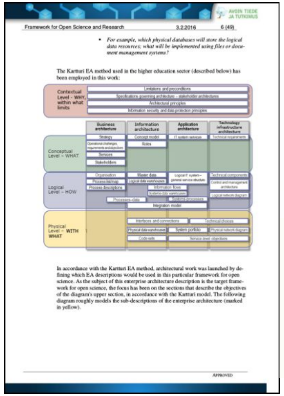
- *For example, which physical databases will store the logical data resources; what will be implemented using files or document management systems?*

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.
Våldsdåd av arbetare.
Återupptagandet av trafiken Åbo—Stockholm.
Pris 25 penni.



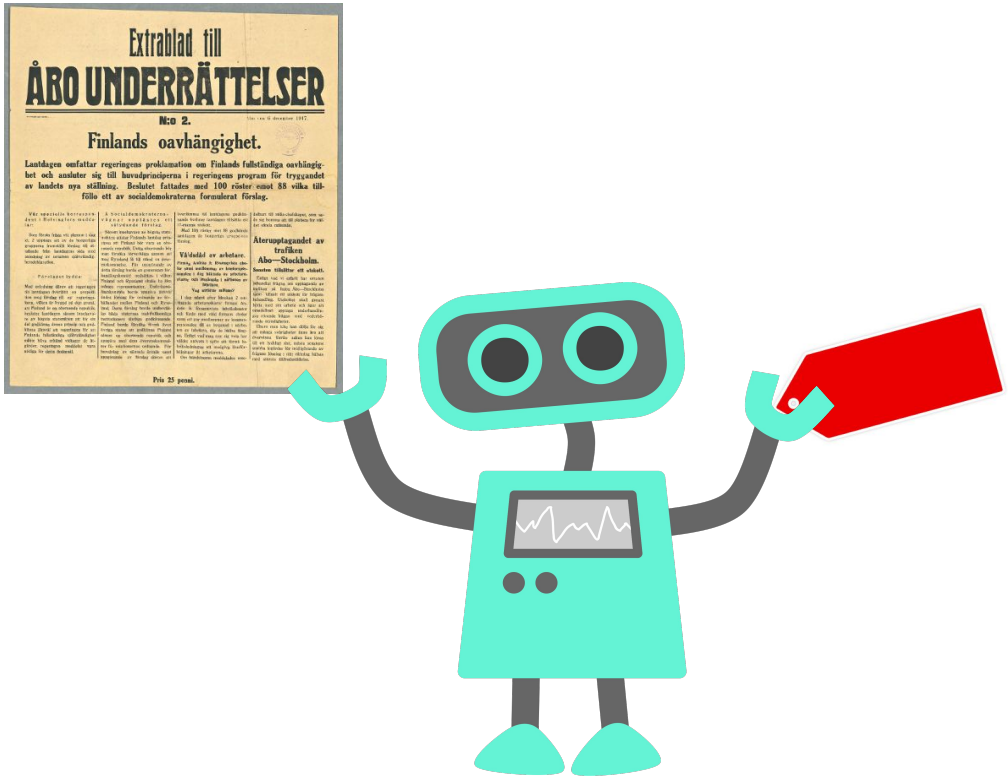

AVOIN TIEDE JA TUTKIMUS
Framework for Open Science and Research 3.2.2016 6 (49)

- *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).

APPROVED



Extrablad till

ÅBO UNDERRÄTTELSER

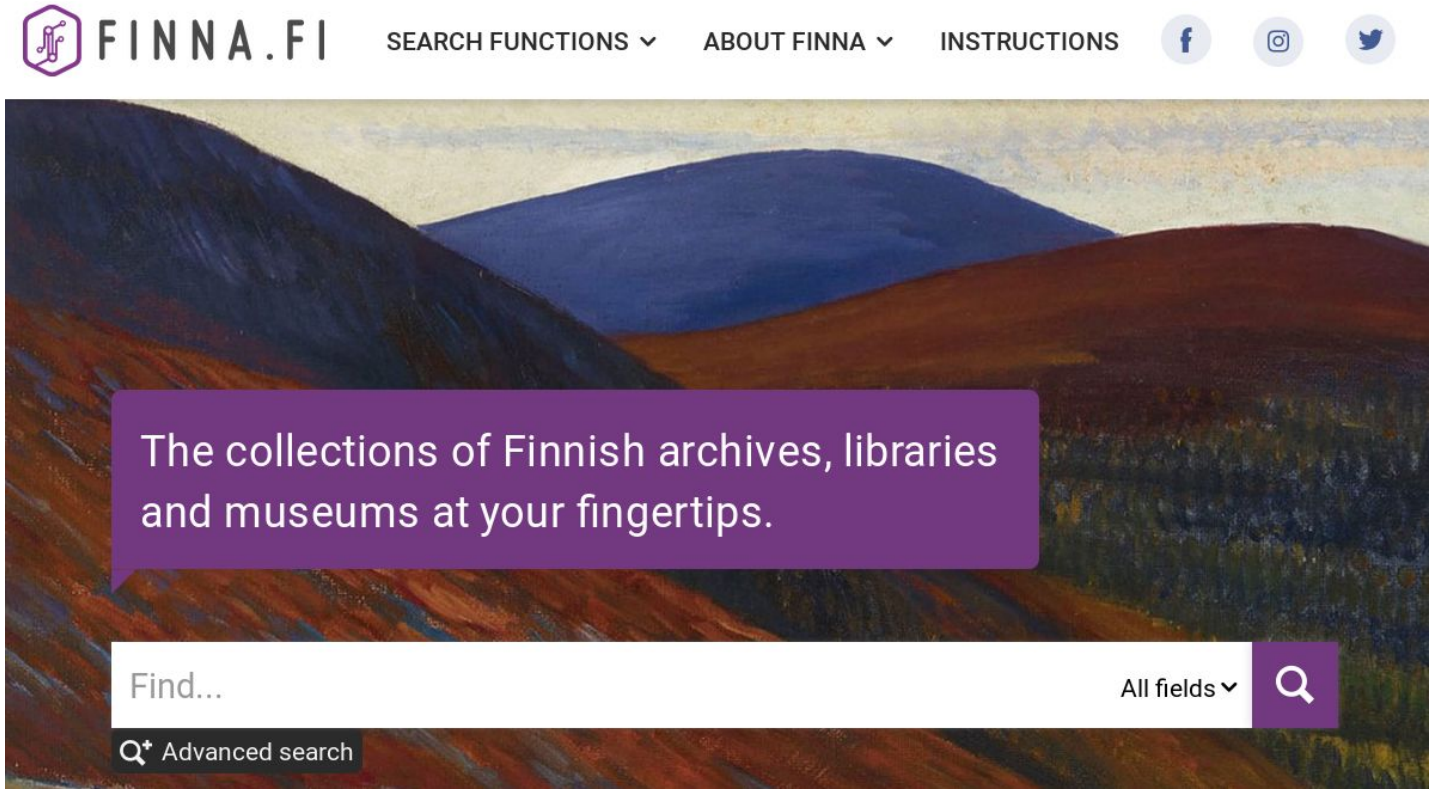
N:o 2.

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.

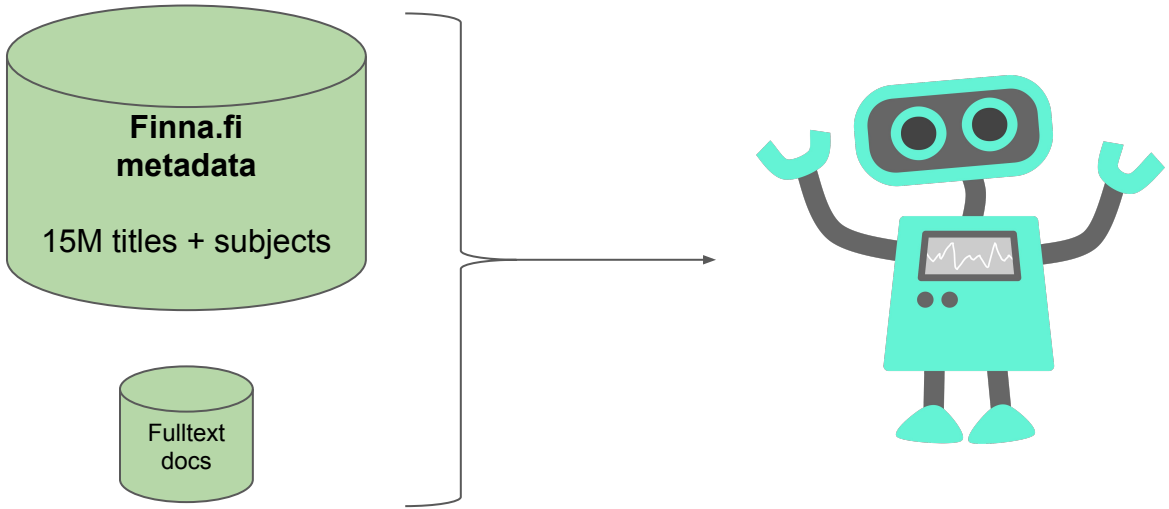
Pris 25 penni.

Idea of Annif

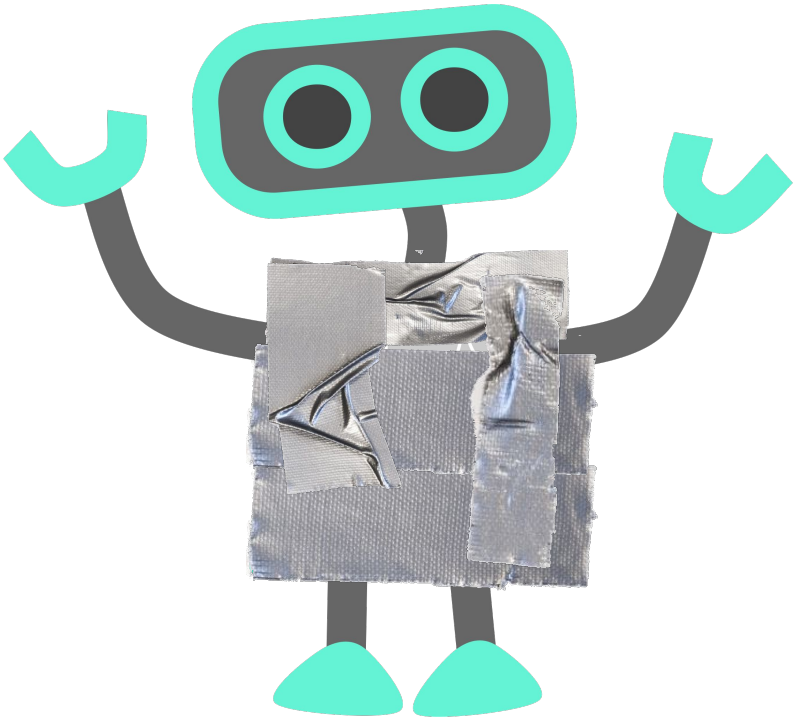


We have **a lot** of LAM metadata, e.g. **15M records** in Finna.fi discovery service

Machine learning using library data



Annif prototype (2017)



Annif prototype vs. new Annif

	Prototype (2017)	**New Annif (2018→)**
architecture	loose collection of scripts	Flask web application
coding style	quick and dirty	solid software engineering
backends	Elasticsearch index	TF-IDF, fastText, Maui ...
language support	Finnish, Swedish, English	any language supported by NLTK
vocabulary support	YSO, GACS ...	YSO, YKL, others coming
REST API	minimal	extended (e.g. list projects)
user interface	web form for testing	http://dev.annif.org
mobile app	HTML/CSS/JS based	native Android app
open source license	CC0	Apache License 2.0

Algorithms for automated subject indexing

Lexical vs. Associative approaches for subject indexing

Lexical approaches

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

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

For more information, see:
Toepfer, M., & Seifert, C. (2018). **Fusion architectures for automatic subject indexing under concept drift: Analysis and empirical results on short texts**. *International Journal on Digital Libraries*. DOI: 10.1007/s00799-018-0240-3

Algorithms used in Annif

Statistical / Associative

- **TF-IDF similarity**
  Baseline bag-of-words similarity measure. Implemented with the Gensim library.

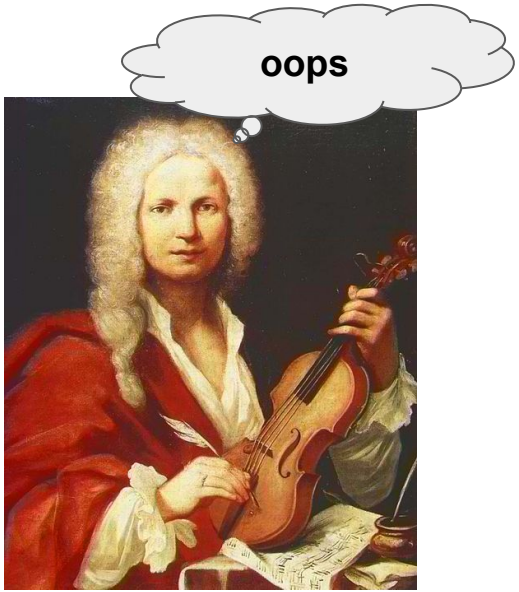
- **fastText** by Facebook Research
  Machine learning algorithm for text classification.
  Uses word embeddings (similar to word2vec) and resembles a neural network architecture.
  Promises to be good for e.g. library classifications (DDC, UDC, YKL…)

Lexical

- **Maui** using MauiService REST API
  MauiService is a microservice wrapper around the Maui automated indexing tool.
  Based on traditional Natural Language Processing techniques - finds terms within text.

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

Algorithms make silly mistakes



Some reasons for mistakes:

- errors and skew in training data
- correlation ≠ causation
- homonyms (e.g. rock)
- misinterpreted names (e.g. Smith, AIDS)
- random noise

In an ensemble, **each algorithm makes different mistakes**

Solution: If we have some more training documents, we can perform **second order learning**!

Isotonic regression, implemented using the Pool Adjacent Violators (**PAV**) algorithm, is a good way of assessing trustworthiness of individual algorithms and turning raw scores into final probability estimates.

Wilbur, W. J., & Kim, W. (2014). **Stochastic Gradient Descent and the Prediction of MeSH for PubMed Records.** *AMIA Annual Symposium proceedings. AMIA Symposium*, *2014*, 1198-207.

Annif Fusion experiment demonstrates PAV

Evaluation of algorithms

Test corpora for evaluating algorithms

Full text documents indexed with YSA/YSO for training and evaluation

1. **Arto:** Articles from Arto database (n=6287)
   Both scientific research papers and less formal publications. Many disciplines.

2. **JYU theses:** Master's and Doctoral theses from University of Jyväskylä (n=7400)
   Long, in-depth scientific documents. Many disciplines.

3. **AskLib:** Question/Answer pairs from an Ask a Librarian service (n=3150)
   Short, informal questions and answers about many different topics.

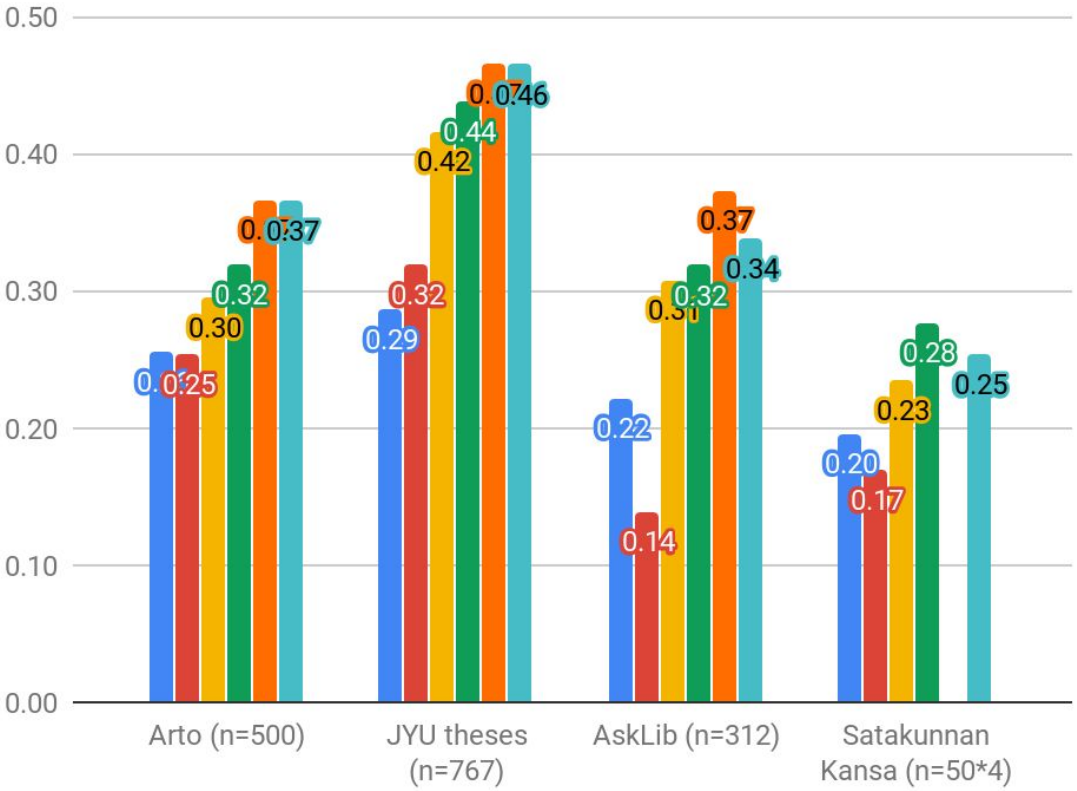
4. **Satakunnan Kansa:** Digital archives of Satakunnan Kansa regional newspaper.
   Over 100k documents, of which 50 have been indexed independently by 4 librarians.

Corpora 1-3 available on GitHub: https://github.com/NatLibFi/Annif-corpora

(for 1-2, only links to PDFs are provided for copyright reasons)

Evaluation of different algorithms in Annif

F1 scores (combination of precision & recall) against gold standard subjects

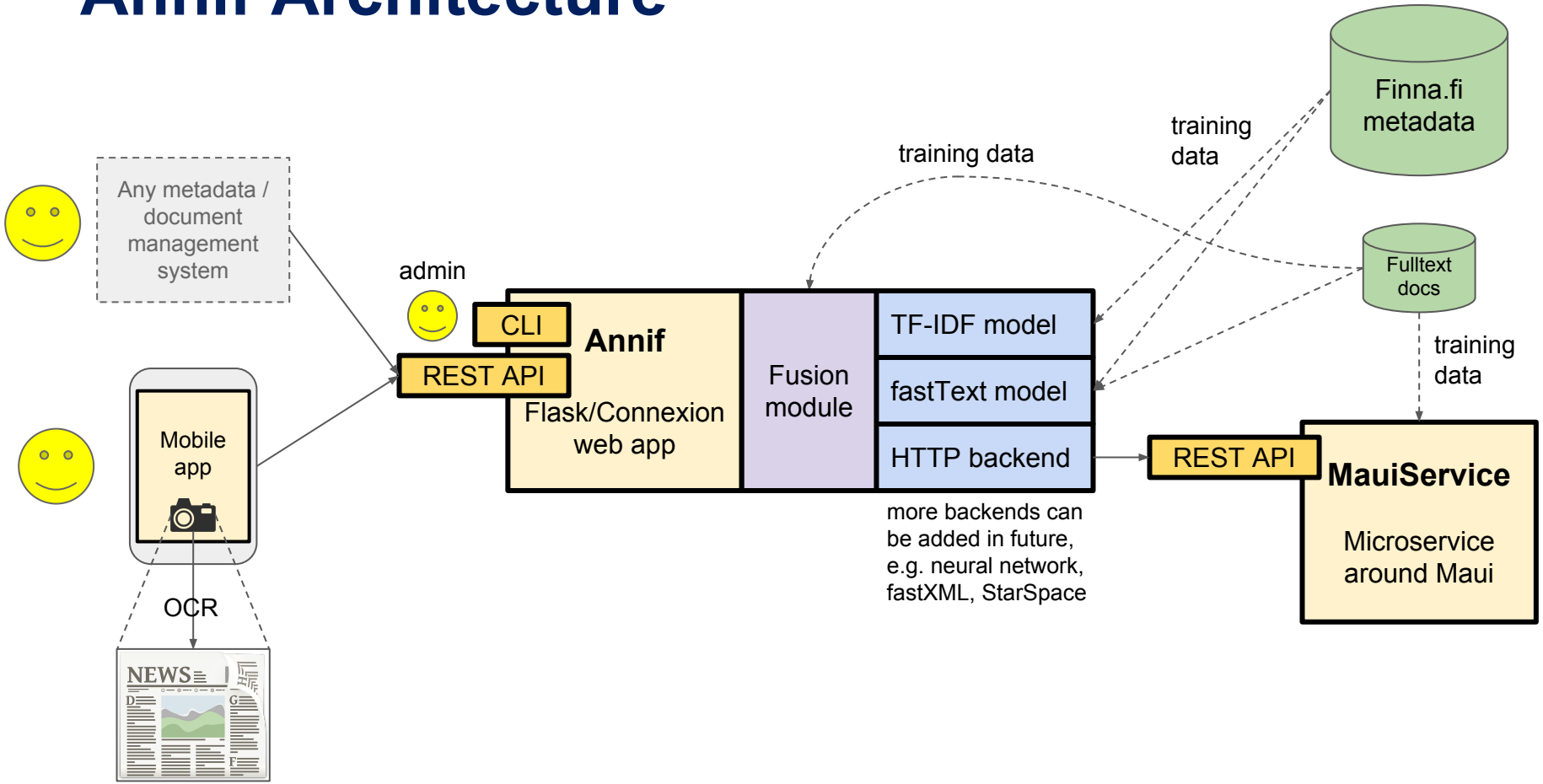


Observations:

1. Of individual algorithms, Maui is the best
2. Ensembles beat individual algorithms
3. PAV ensembles can be better than a simple ensemble (but not always)

Software architecture

Annif Architecture



Form for testing at annif.org

YSO model
trained on Finna data

Try Annif!

Text to analyze:

SWIB conference (Semantic Web in Libraries) is an annual conference, being held for the 10th time, focusing 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 all 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.

Project (vocabulary and language):

YSO ensemble English

Analyze

Results

- open data
- semantic web
- linked open data
- semantics
- data
- applied sciences
- librarians
- data science
- libraries
- Works

Form for testing at annif.org

Wikidata model
trained on Wikipedia

Top 50k items by # of sitelinks

Try Annif!

Text to analyze:

SWIB conference (Semantic Web in Libraries) is an annual conference, being held for the 10th time, focusing 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 all 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.

Project (vocabulary and language):

Wikidata TF-IDF English

Analyze

Results

- Lod
- big data
- librarian
- library science
- data mining
- data
- Semantic Web
- library
- Library of Congress
- data structure

Command line interface

Load a vocabulary to be used by one or more models:

```
$ annif loadvoc tfidf-en yso-en.tsv
```

Train a model:

```
$ annif train tfidf-en yso-finna-en.tsv.gz
```

Analyze a document:

```
$ annif analyze tfidf-en <berries.txt
<http://www.yso.fi/onto/yso/p772>    strawberry           0.39644203283656165
<http://www.yso.fi/onto/yso/p18109>  wild strawberry      0.37539359094384245
<http://www.yso.fi/onto/yso/p25548>  stolons              0.3261554545369906
<http://www.yso.fi/onto/yso/p6749>   berry cultivation    0.2394291077460799
<http://www.yso.fi/onto/yso/p10631>  questionnaire survey 0.22714475653823335
<http://www.yso.fi/onto/yso/p6821>   farms                0.21725243067995587
<http://www.yso.fi/onto/yso/p3294>   customers            0.216395821347059
<http://www.yso.fi/onto/yso/p1834>   work motivation      0.21612376226244975
<http://www.yso.fi/onto/yso/p8531>   customership         0.21536113638508098
<http://www.yso.fi/onto/yso/p19047>  corporate clients    0.21412270159920782
```

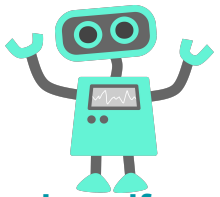
Evaluate a model using several measures (e.g. recall, precision, F1 score, NDCG):

```
$ annif eval tfidf-en directory-with-gold-standard-docs/
```

REST API access example

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

Analyze this!



api.annif.org

results=[

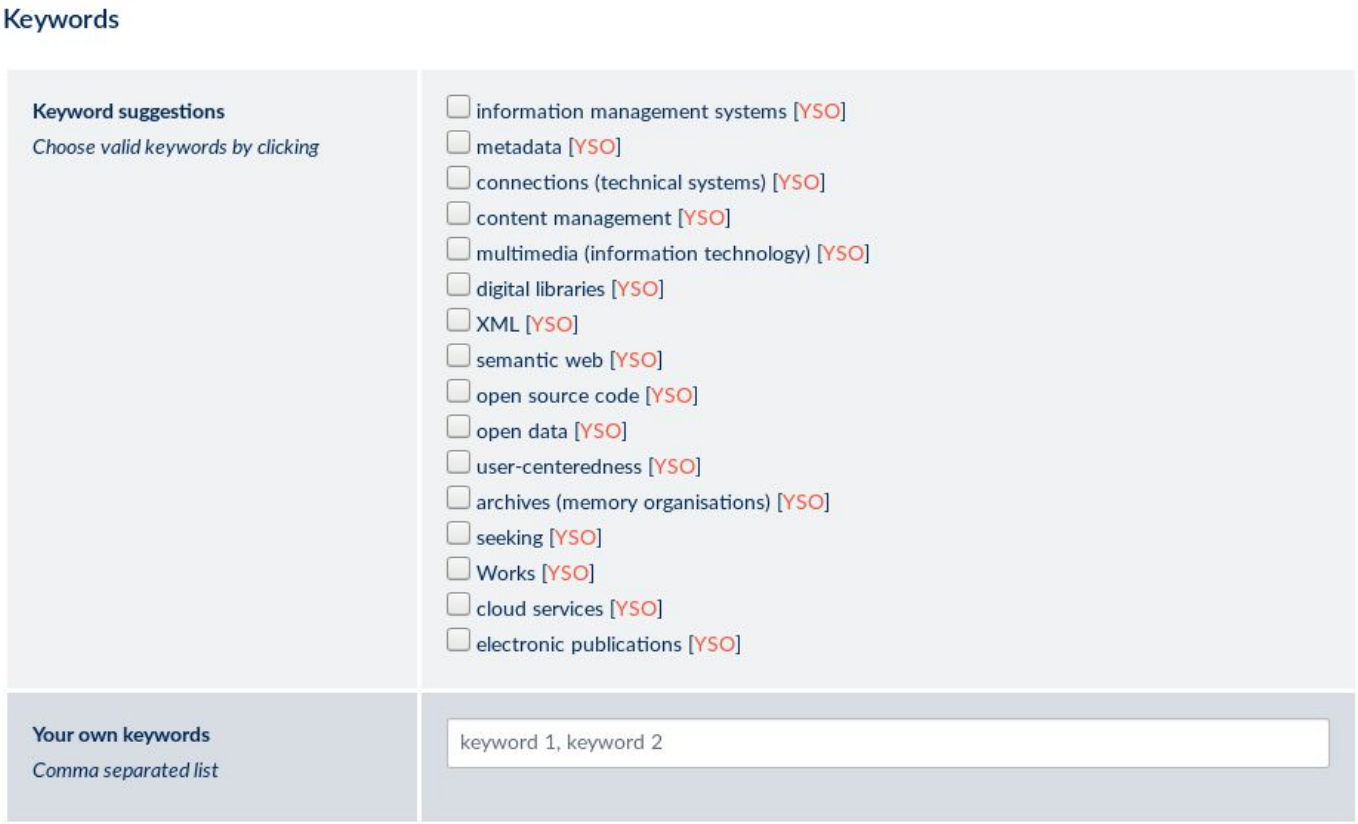
{uri=”<http://www.yso.fi/onto/yso/p2228>”, score=0.2595, label=”**red fox**”},
{uri=”<http://www.yso.fi/onto/yso/p5319>”, score=0.2039, label=”**dog**”},
{uri=”<http://www.yso.fi/onto/yso/p8122>”, score=0.1946, label=”**laziness**”},
{uri=”<http://www.yso.fi/onto/yso/p25726>”, score=0.1285, label=”**brown**”},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

]

What can you do with Annif?

JYX repository, University of Jyväskylä

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



Implemented using DSpace & GLAMpipe by Ari Häyrinen

Indexing Wikipedia by topics

Finnish Wikipedia has 410 000 articles (620 MB as raw text)
Automated subject indexing took 7 hours on a laptop, using the Annif prototype
1-3 topics per article (average ~2)

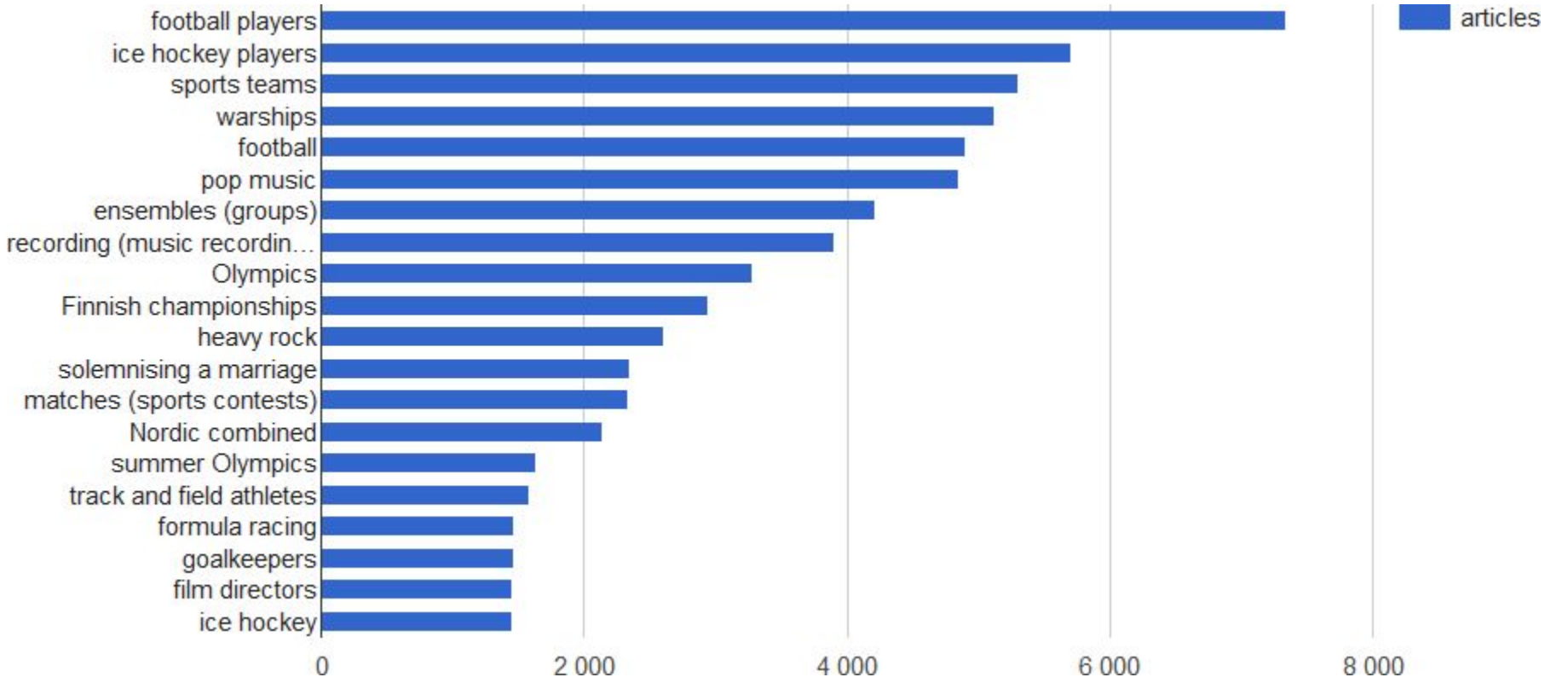
Indexing Wikipedia by topics

Finnish Wikipedia has 410 000 articles (620 MB as raw text)
Automated subject indexing took 7 hours on a laptop
1-3 topics per article (average ~2)

Examples: (random sample)

Wikipedia article	YSO topics
Ahvenuslammi (Urjala)	shores
Brasilian Grand Prix 2016	race drivers, formula racing, karting
Guy Topelius	folk poetry researcher, saccharin
HMS Laforey	warships
Liigacup	football, football players
Pää Kii	ensembles (groups), pop music
RT-21M Pioneer	missiles
Runoja	pop music, recording (music recordings), compositions (music)
Sjur Røthe	skiers, skiing, Nordic combined
Veikko Lavi	lyricists, comic songs

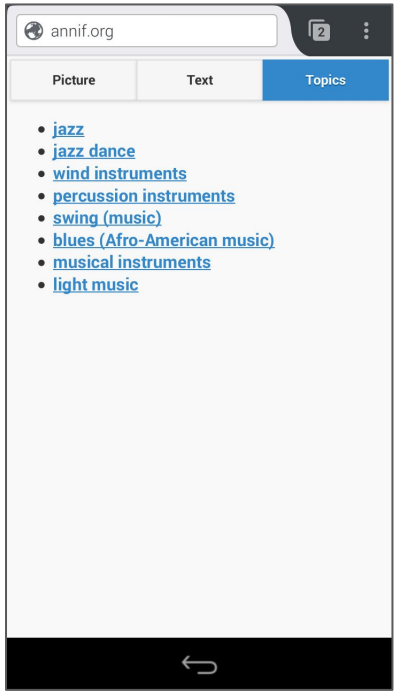
Most common topics in Finnish Wikipedia



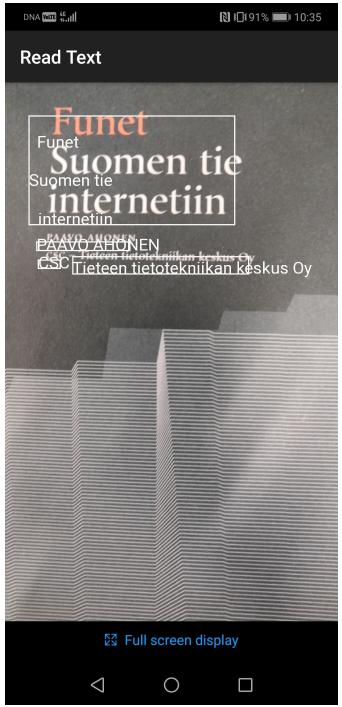
Most common topics in Finnish Wikipedia

Image credits:
Petteri Lehtonen [CC BY-SA 3.0]
Hockeybroad/Cheryl Adams [CC BY-SA 3.0]
Tomisti [CC BY-SA 3.0]
Tuomas Vitikainen [CC BY-SA 3.0]

Mobile apps

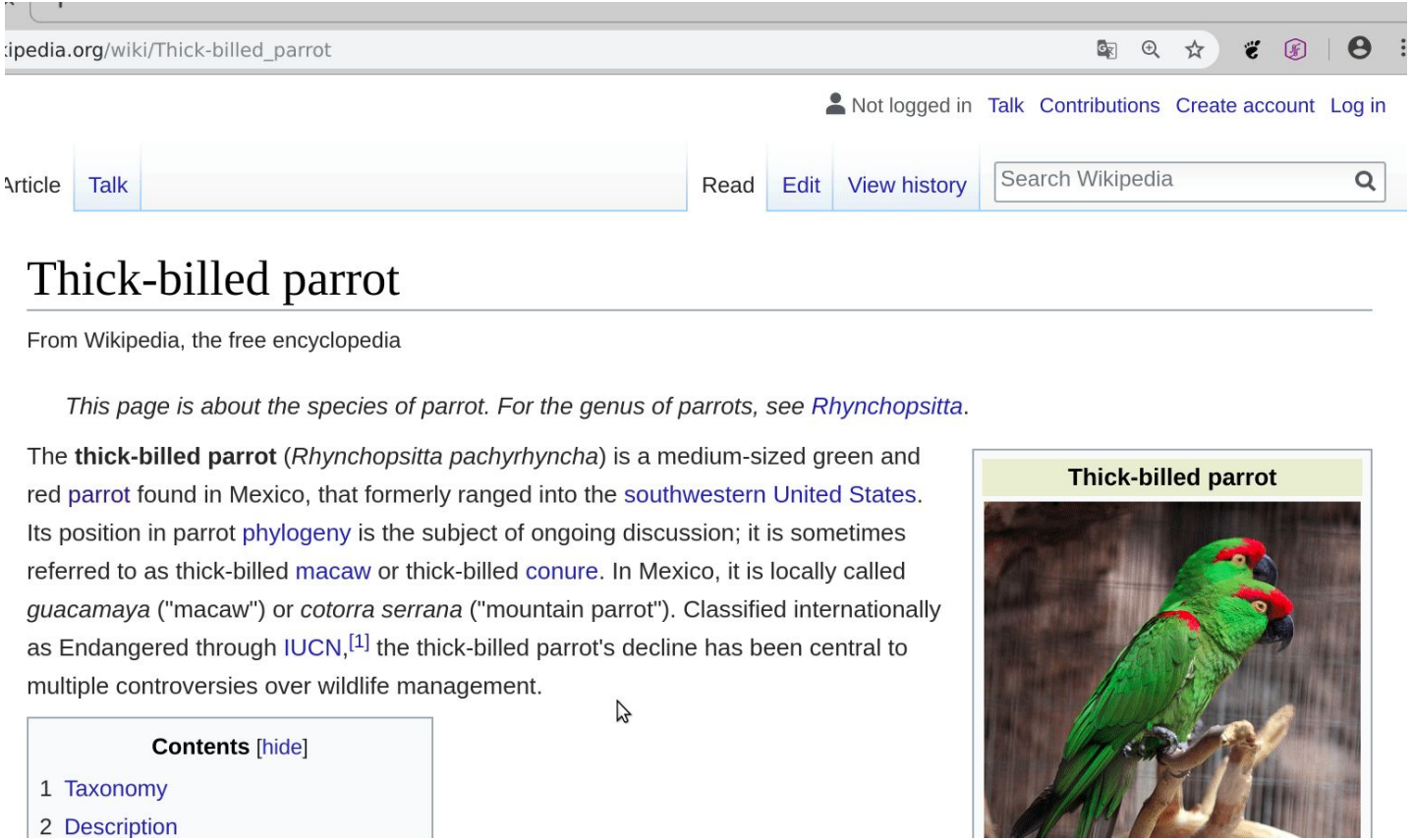


Prototype web app,
ocr.space cloud OCR
m.annif.org



Prototype Android app with OCR on the device
(by Okko Vainonen)

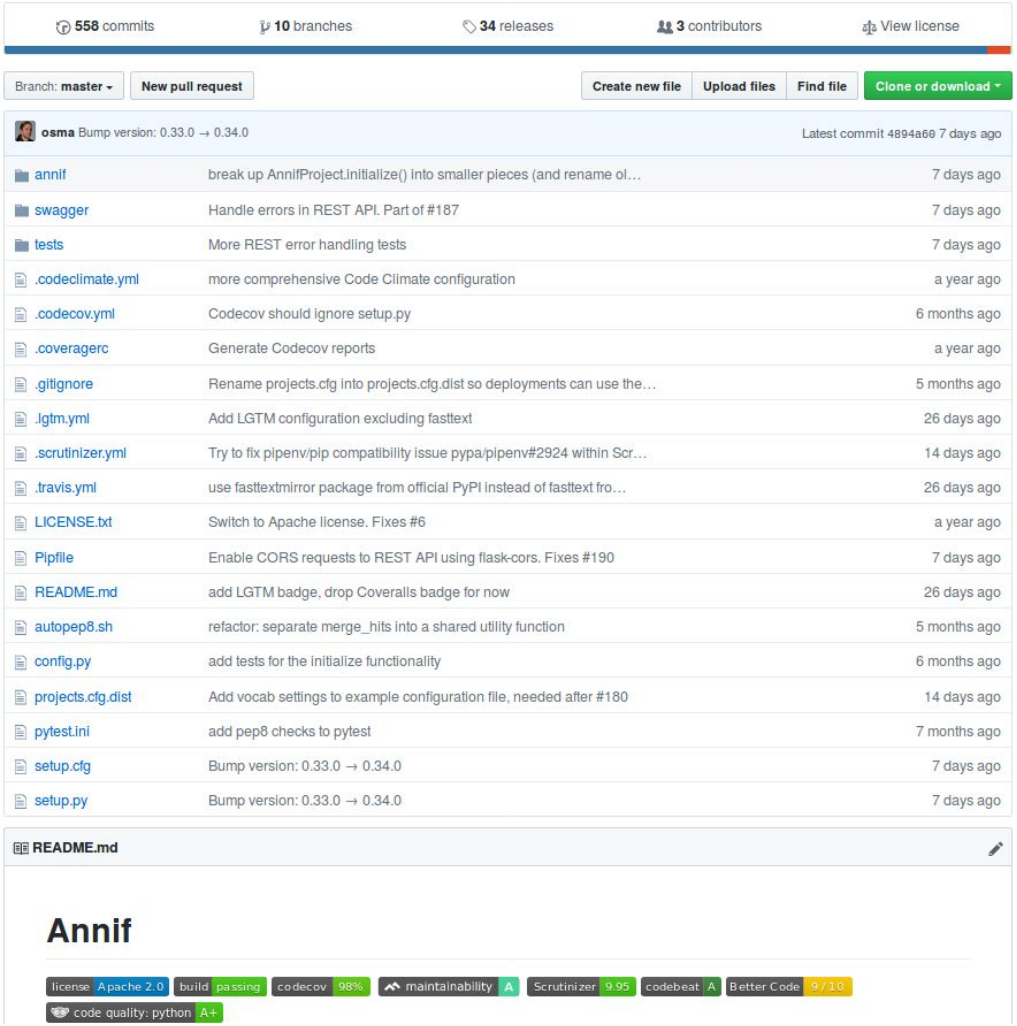
Finna Recommends Chrome browser extension



Analyzes selected text from any web page using Annif API and **recommends books** from Finna.fi

Created during **WIDE hackathon** by
Yazan Alhalabi
Samuel Akangbe
Steven Nebo

Getting Annif

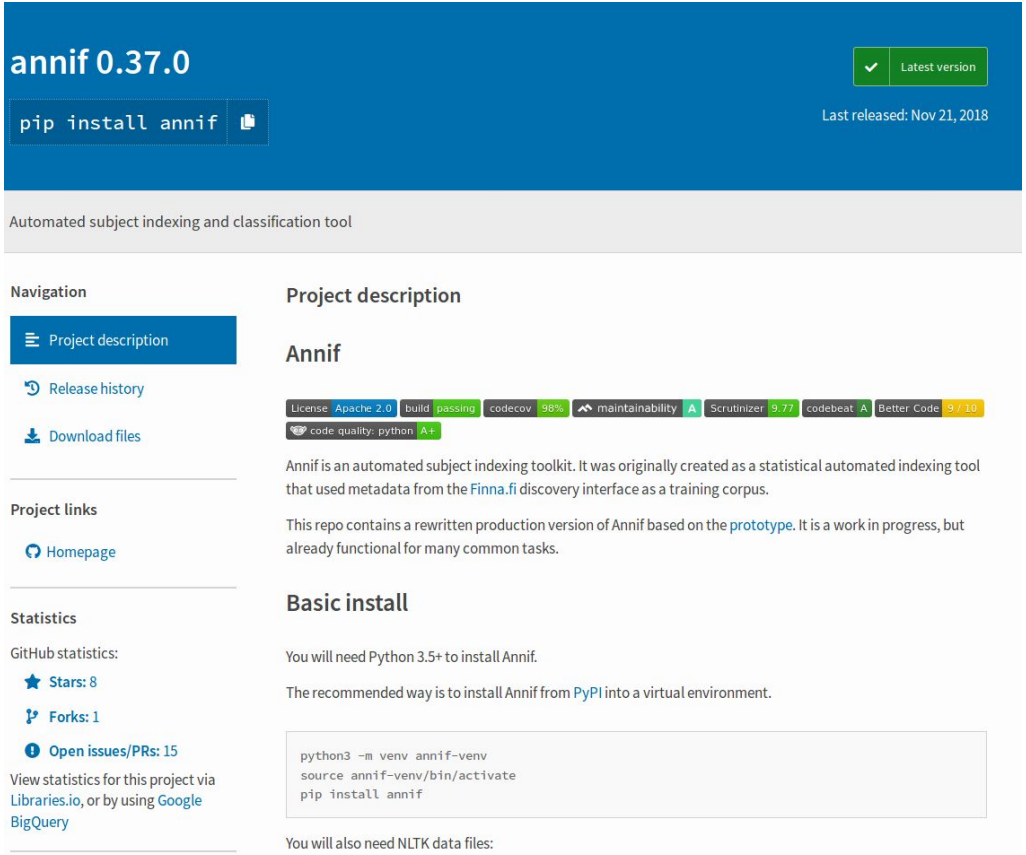


Annif on GitHub

Python 3.5+ code base
Apache License 2.0

Fully unit tested (98% coverage)
PEP8 style guide compliant
Usage documentation in the wiki

https://github.com/NatLibFi/Annif



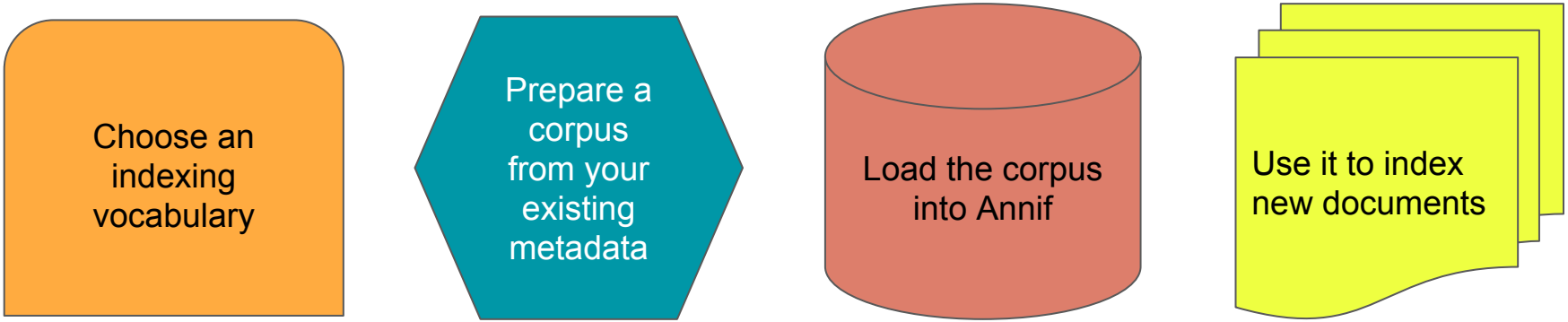
Annif on PyPI

Installing into a virtualenv:

```
pip install annif
```

https://pypi.org/project/annif/

Apply Annif on your own data!



Community group on DIY automated subject indexing?

To discuss applications, algorithms, API standards, corpora, formats etc.

Contact me if interested!

Thank you!
Questions?

osma.suominen@helsinki.fi - @OsmaSuominen

Website: http://annif.org
API: http://api.annif.org

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