Automated Subject Indexing and Classification using Annif

Osma Suominen

HELDIG Summit, 23 October 2018

Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

Åbo den 6 december 1917.

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 i Helsingfors meddelar:

[illegible]

Förslaget lydde:

[illegible]

Å Socialdemokraternas vägnar uppläses ett avvikande förslag.

[illegible]

Våldsdåd av arbetare.

[illegible]

Återupptagandet av trafiken Åbo—Stockholm.

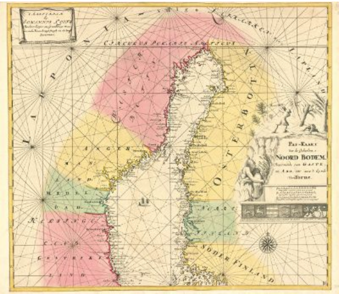
Senaten tillsätter ett utskott.

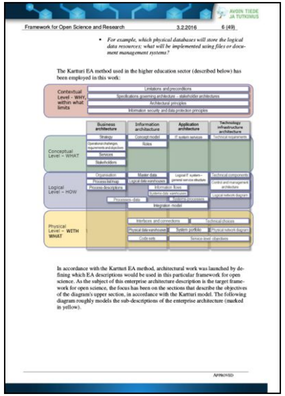
[illegible]

Pris 25 penni.

VENNY SOLDAN-BROFELDT.

Unikko. (Papaver orientale.)





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:

Contextual Level – WHY, within what limits

Limitations and preconditions

Specifications governing architecture – stakeholder architectures

Architectural principles

Information security and data protection principles

Business architecture

Information architecture

Application architecture

Technology infrastructure architecture

Conceptual Level – WHAT

Strategy

Operational challenges, requirements and objectives

Services

Stakeholders

Concept model

Roles

IT system services

Technical requirements

Logical Level – HOW

Organisation

Process list/map

Process descriptions

Master data

Logical data warehouses

Logical IT system – general service structure

Technical components

Control and management architecture

Information flows

Systems-data warehouses

Logical network diagram

Processes–data

Systems processes

Integration model

Physical Level – WITH WHAT

Interfaces and connections

Technical choices

Physical data warehouses

System portfolio

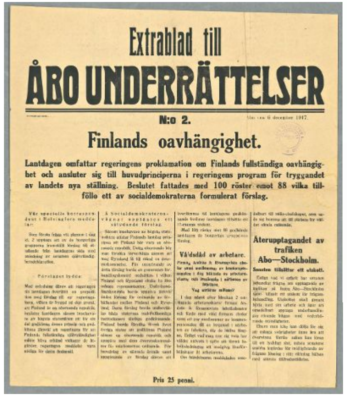
Physical network diagram

Code sets

Service level objectives

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. Åbo den 6 december 1917.

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 i Helsingfors meddelar:

[illegible]

Förslaget lydde:

[illegible]

Å Socialdemokraternas vägnar uppläses ett självdande förslag.

[illegible]

Våldsdåd av arbetare.

[illegible]

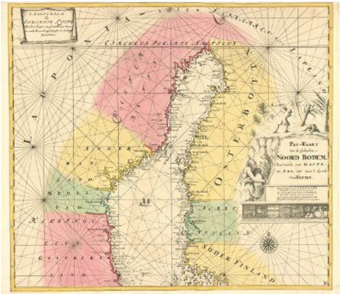
Återupptagandet av trafiken Åbo—Stockholm.

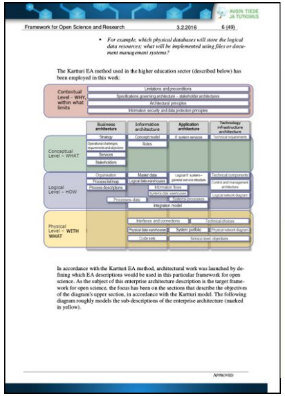
Senaten tillsätter ett utskott.

[illegible]

Pris 25 penni.

VENNY SOLDAN-BROFELDT. Unikko. (Papaver orientale.)





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?*

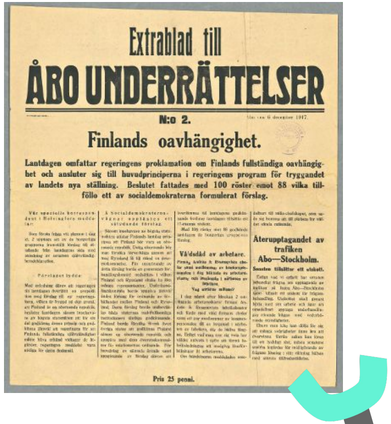
The Kartturi EA method used in the higher education sector (described below) has been employed in this work:

Contextual Level - WHY, within what limits
Limitations and preconditions
Specifications governing architecture – stakeholder architectures
Architectural principles
Information security and data protection principles

	Business architecture	Information architecture	Application architecture	Technology infrastructure architecture
Conceptual Level – WHAT	Strategy Operational challenges, requirements and objectives Services Stakeholders	Concept model Roles	IT system services	Technical requirements
Logical Level – HOW	Organisation Process list/map Process descriptions Processes-data	Master data Logical data warehouses Information flows Systems-data warehouses Integration model	Logical IT system – general service structure Systems processes	Technical components Control and management architecture Logical network diagram
Physical Level – WITH WHAT		Interfaces and connections Physical data warehouses Code sets	System portfolio Service level objectives	Technical choices Physical network diagram

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.

[illegible] den 6 december 1917.

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 tillfölle ett av socialdemokraterna formulerat förslag.

Vår speciella korrespondent i Helsingfors meddelar:

[illegible]

Förslaget lyder:

[illegible]

Å Socialdemokraternas vägnar uppläses ett sålydande förslag.

[illegible]

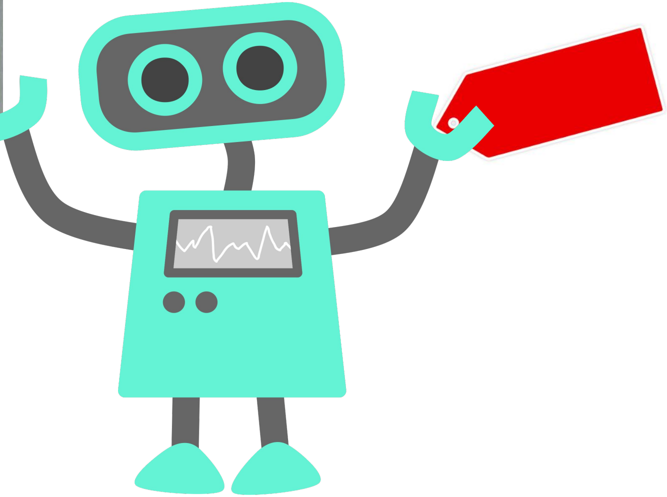
Våldsdåd av arbetare.

[illegible]

Återupptagandet av trafiken Åbo—Stockholm.

[illegible]

Pris 25 penni.

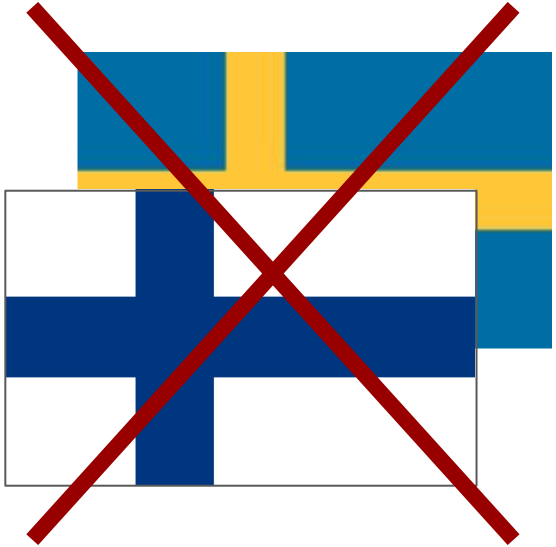


DBpedia Spotlight

Leiki

OPEN
CALAIS
THOMSON REUTERS

apache
stanbol™



YSA
Allärs

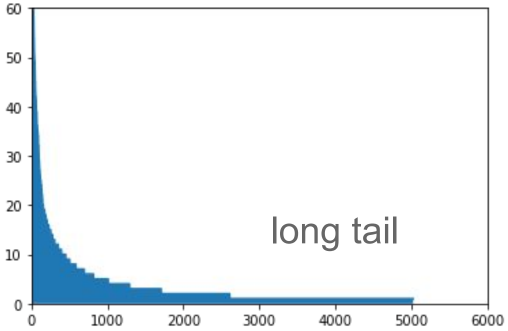
YSO
KOKO

€£$

Subject indexing is a hard problem

for **humans:**

- **Subjectivity**: when two people index the same document, only ~⅓ of the subjects are the same
- **Many concepts**: tens of thousands of concepts to pick from
- **Vocabulary changes:** new concepts are added, existing ones are renamed and redefined

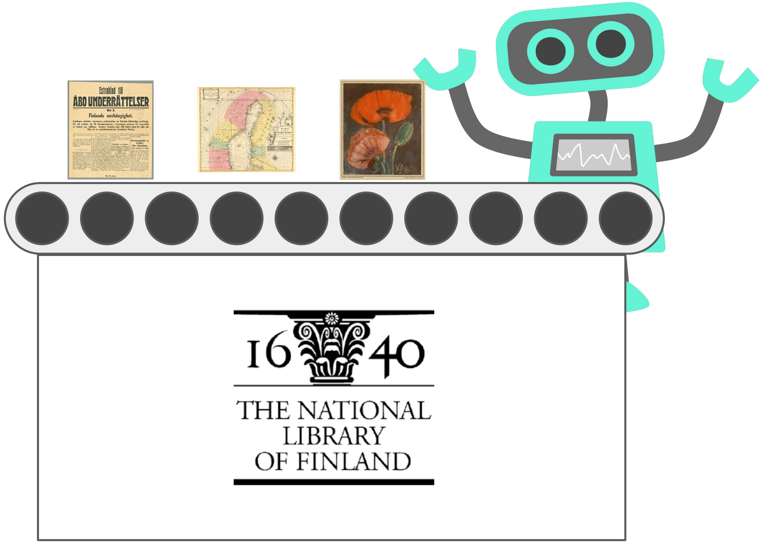


for **machines:**

- **Long tail phenomenon:** even with large amounts of training data, most subjects are only used a small number of times
- **Many concepts**: requires complex models that are computationally intensive
- **Difficult to evaluate**: hard to tell “somewhat bad” answers from really wrong ones without human evaluation
- **Vocabulary changes**: models must be retrained

Approach

Automating our own processes

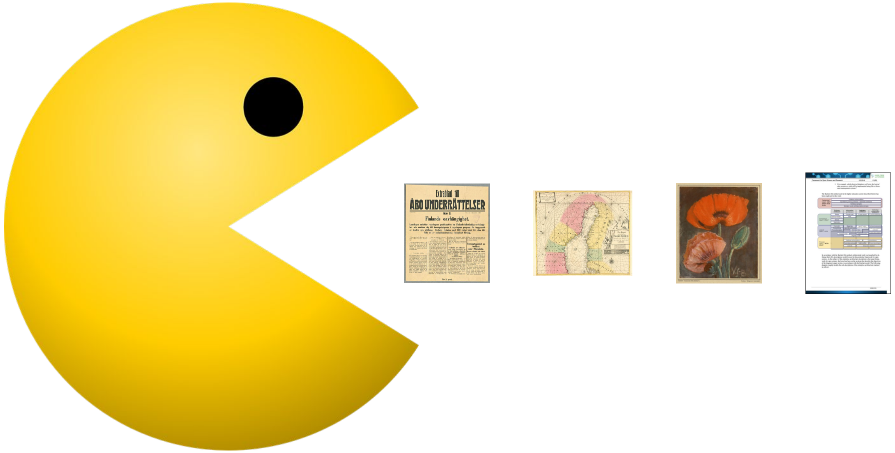


vs.

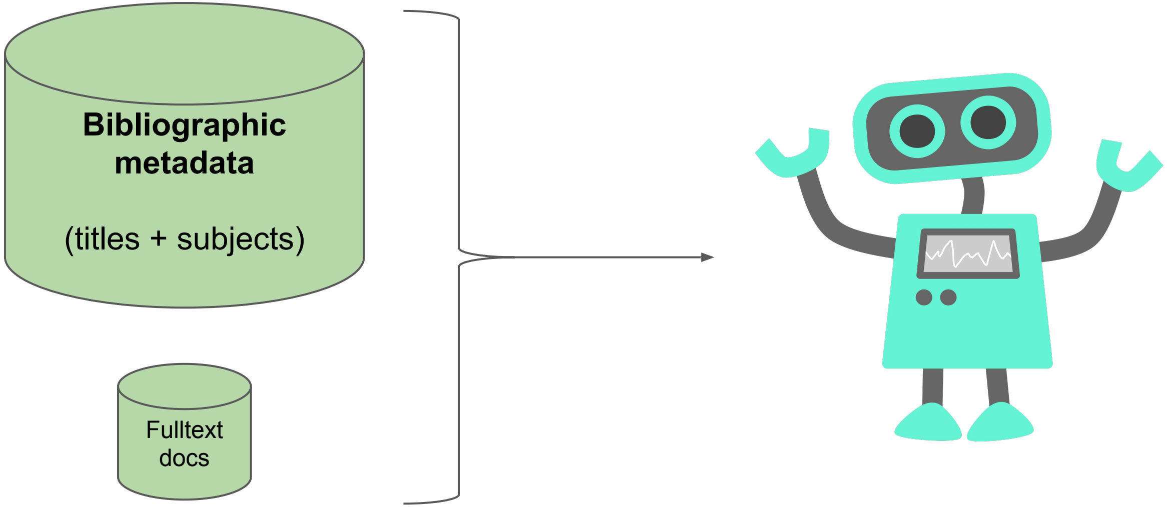
Creating generic tools for many contexts

Enter **Annif**

Feed your subject indexing robot with bibliographic metadata!



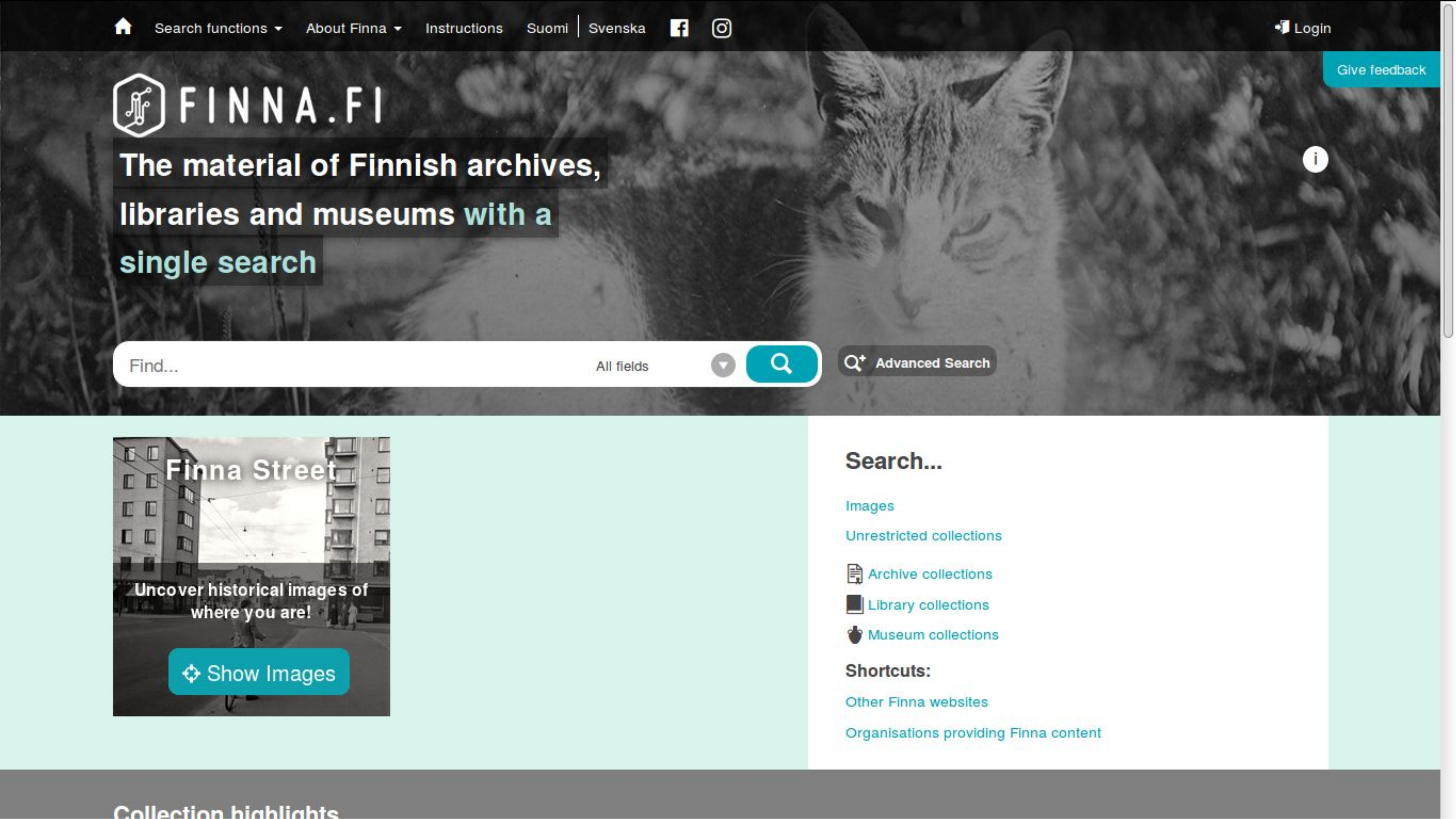
Machine learning requires training data



Bibliographic metadata

(titles + subjects)

Fulltext docs



The material of Finnish archives, libraries and museums with a single search

Finna Street

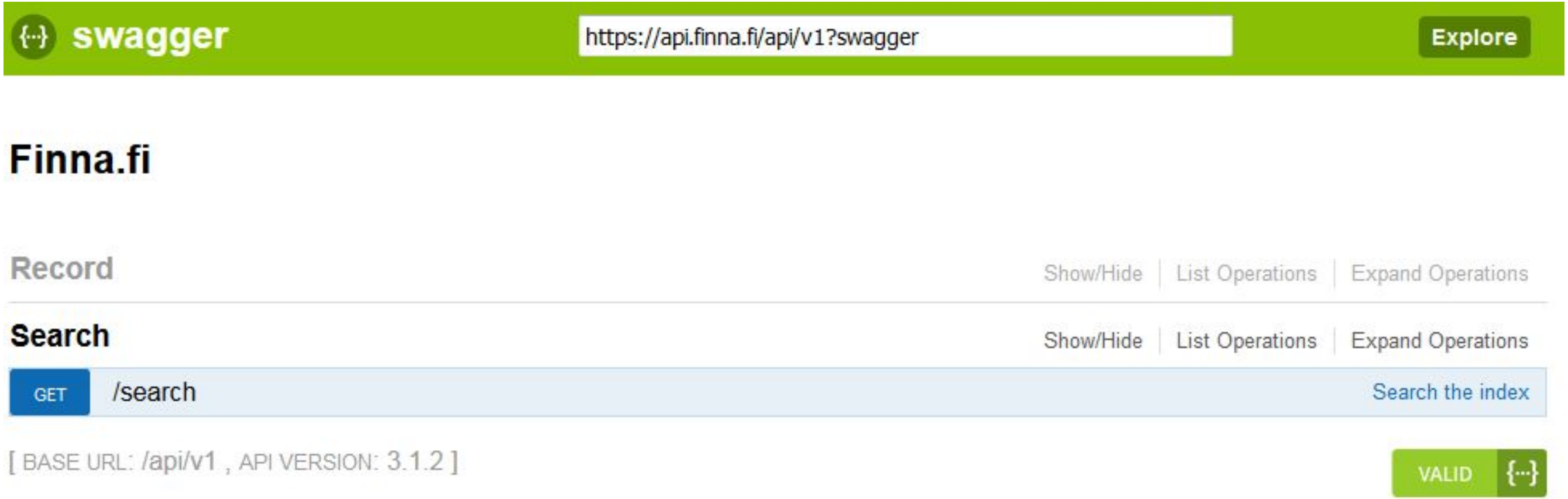
Uncover historical images of where you are!

Search...

Shortcuts:

Collection highlights

Finna API



https://api.finna.fi/api/v1?swagger

Finna.fi

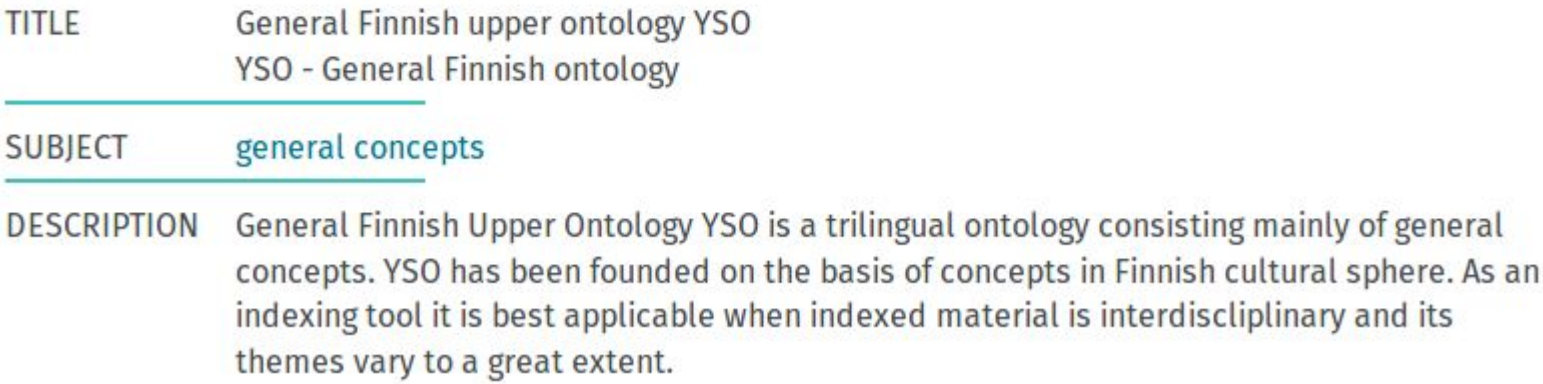
Record

Search

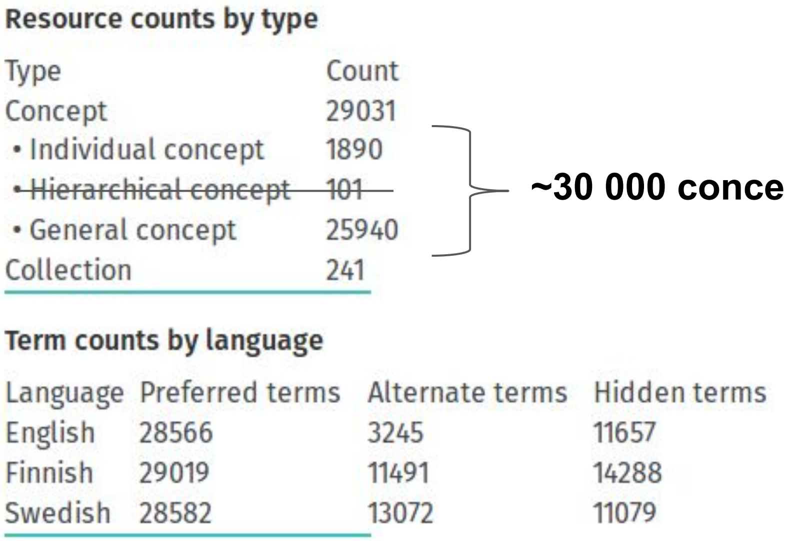
GET /search Search the index

[BASE URL: /api/v1 , API VERSION: 3.1.2]

All Finna metadata is PUBLIC DOMAIN !



TITLE	General Finnish upper ontology YSO YSO - General Finnish ontology
SUBJECT	general concepts
DESCRIPTION	General Finnish Upper Ontology YSO is a trilingual ontology consisting mainly of general concepts. YSO has been founded on the basis of concepts in Finnish cultural sphere. As an indexing tool it is best applicable when indexed material is interdiscliplinary and its themes vary to a great extent.



Resource counts by type

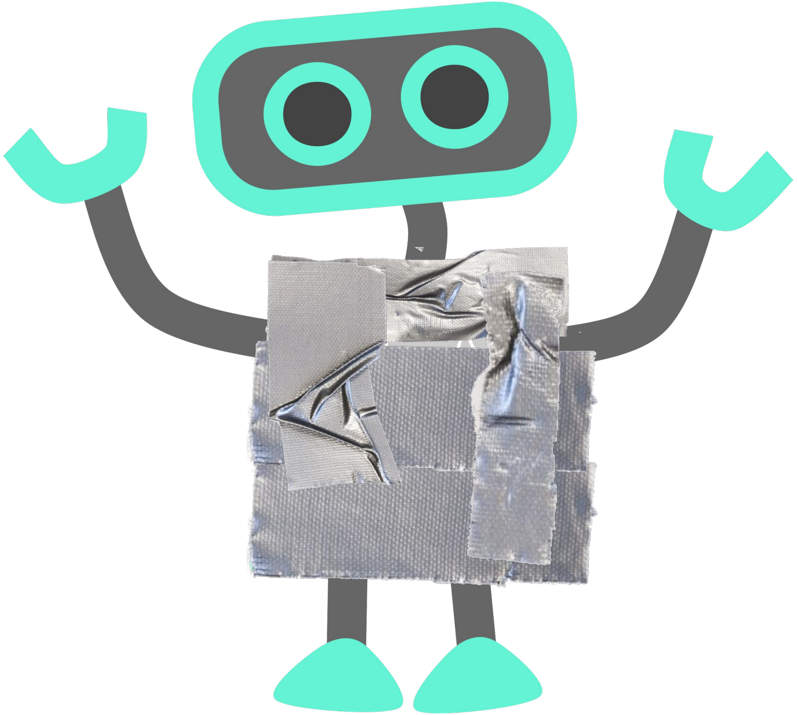
Type	Count
Concept	29031
• Individual concept	1890
~~• Hierarchical concept~~	~~101~~
• General concept	25940
Collection	241

~30 000 concepts that can be used for subject indexing

Term counts by language

Language	Preferred terms	Alternate terms	Hidden terms
English	28566	3245	11657
Finnish	29019	11491	14288
Swedish	28582	13072	11079

Annif prototype (2017)



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)

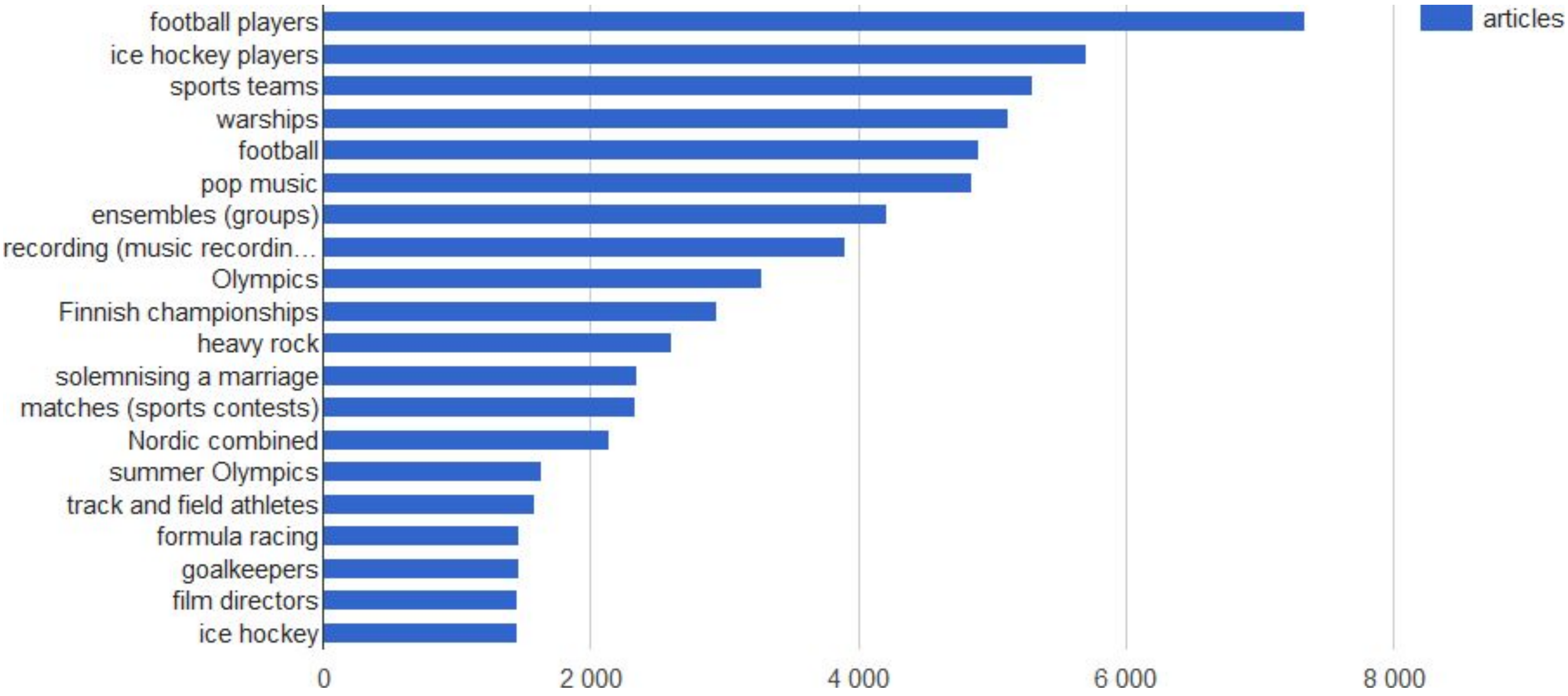
Automated subject indexing took 7 hours on a laptop

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

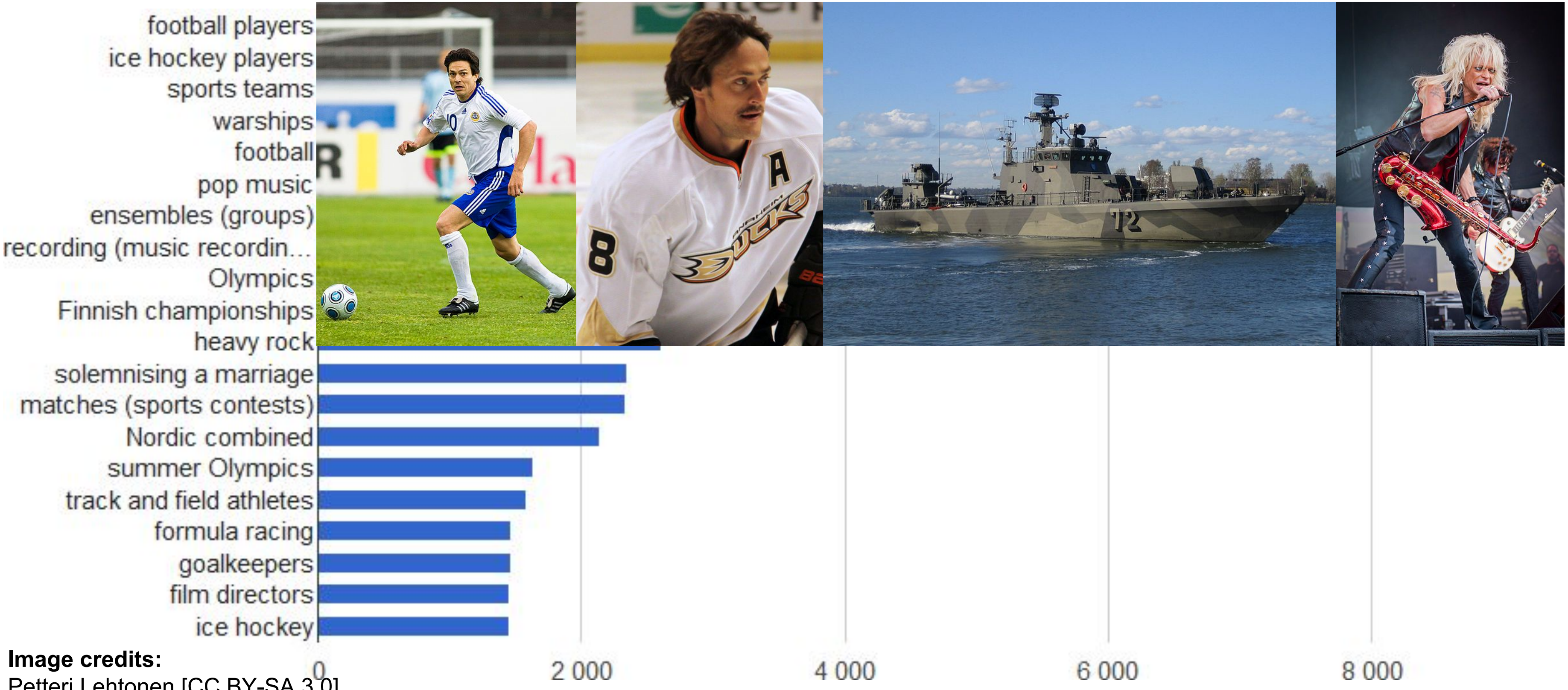
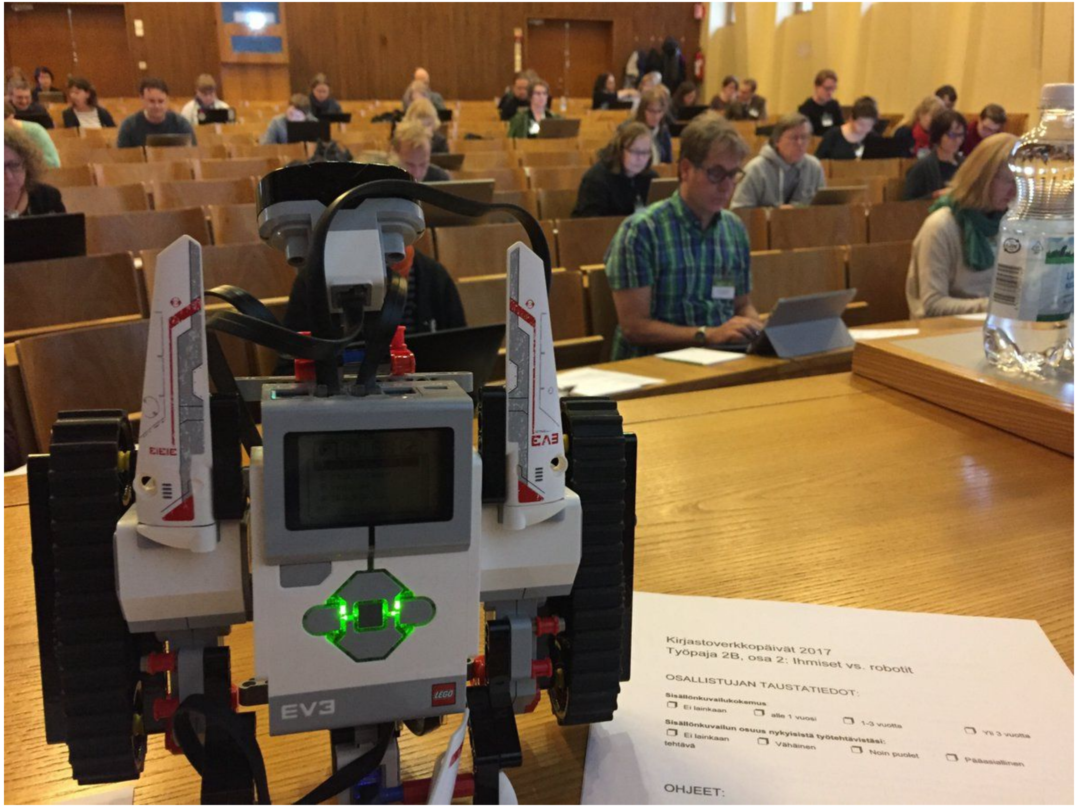


Image credits:
Petteri Lehtonen [CC BY-SA 3.0]
Hockeybroad/Cheryl Adams [CC BY-SA 3.0]
Tomisti [CC BY-SA 3.0]
Tuomas Vitikainen [CC BY-SA 3.0]

People vs. Robots Workshop


EV3
LEGO
Kirjastoverkkopäivät 2017
Työpaja 2B, osa 2: Ihmiset vs. robotit
OSALLISTUJAN TAUSTATIEDOT:
Sisällönkuvailukokemus
□ Ei lainkaan □ alle 1 vuosi □ 1-3 vuotta □ Yli 3 vuotta
Sisällönkuvailun osuus nykyisistä työtehtävistäsi:
□ Ei lainkaan □ Vähäinen □ Noin puolet □ Pääasiallinen tehtävä
OHJEET:

20 documents
40 librarians
45 minutes

...

225 indexing results

- 11 per document
- 5.5 per person

Average similarity of subject sets

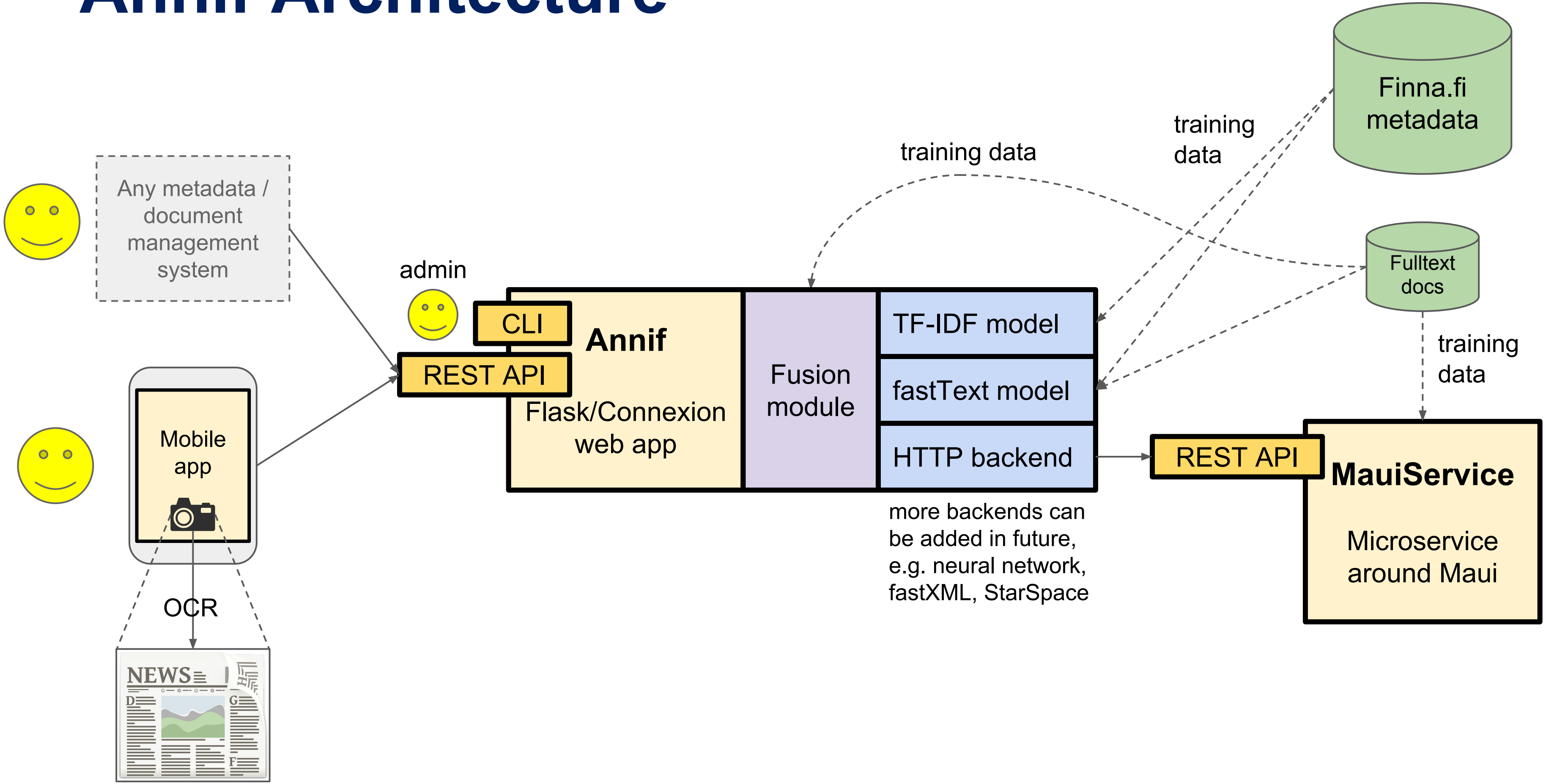
33.39 %

Using Rolling similarity, a.k.a. F1 score, to compare subject sets

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

Annif Architecture



more backends can be added in future, e.g. neural network, fastXML, StarSpace

Backends / Algorithms

- **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…)

- **HTTP backend** for accessing 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.

Backend configuration

Backends may be used alone, or in combinations (ensembles)

Current challenge: Which fusion method works best for combining results from multiple backends?
An experiment testing different fusion methods

Command line interface

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

```
$ annif loadvoc yso-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/
```

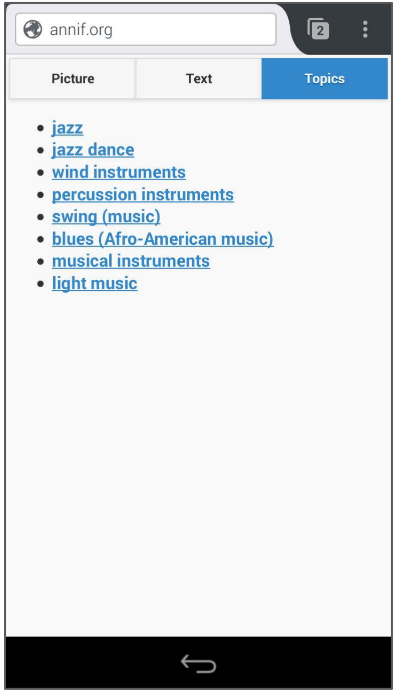
REST API

Main operations:

GET /projects/	list available projects
GET /projects/<project_id>	show information about a project
POST /projects/<project_id>/analyze	analyze text and return subjects
POST /projects/<project_id>/explain	analyze text and return subjects, with explanations indicating why they were chosen
POST /projects/<project_id>/train	train the model by giving a document and gold standard subjects

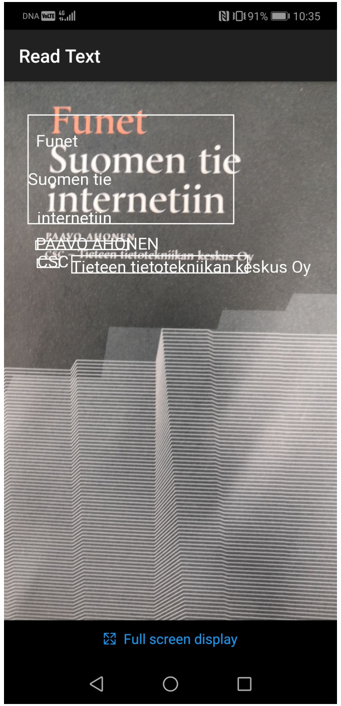
Defined using a Swagger / OpenAPI specification

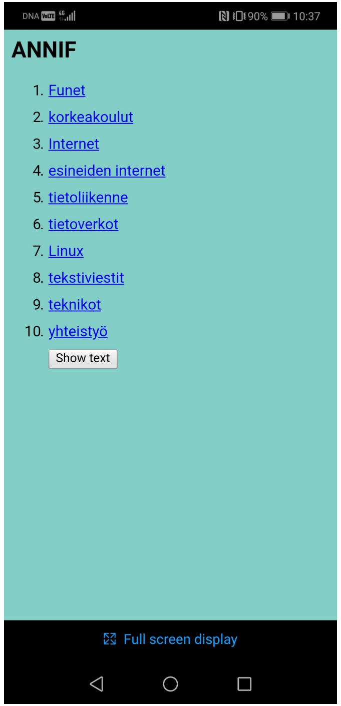
Mobile apps



- jazz
- jazz dance
- wind instruments
- percussion instruments
- swing (music)
- blues (Afro-American music)
- musical instruments
- light music

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





ANNIF

1. Funet
2. korkeakoulut
3. Internet
4. esineiden internet
5. tietoliikenne
6. tietoverkot
7. Linux
8. tekstiviestit
9. teknikot
10. yhteistyö

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

Test corpora

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

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

- Master’s and Doctoral theses from Jyväskylä University (n=7400)
  Long, in-depth scientific documents. Many disciplines.

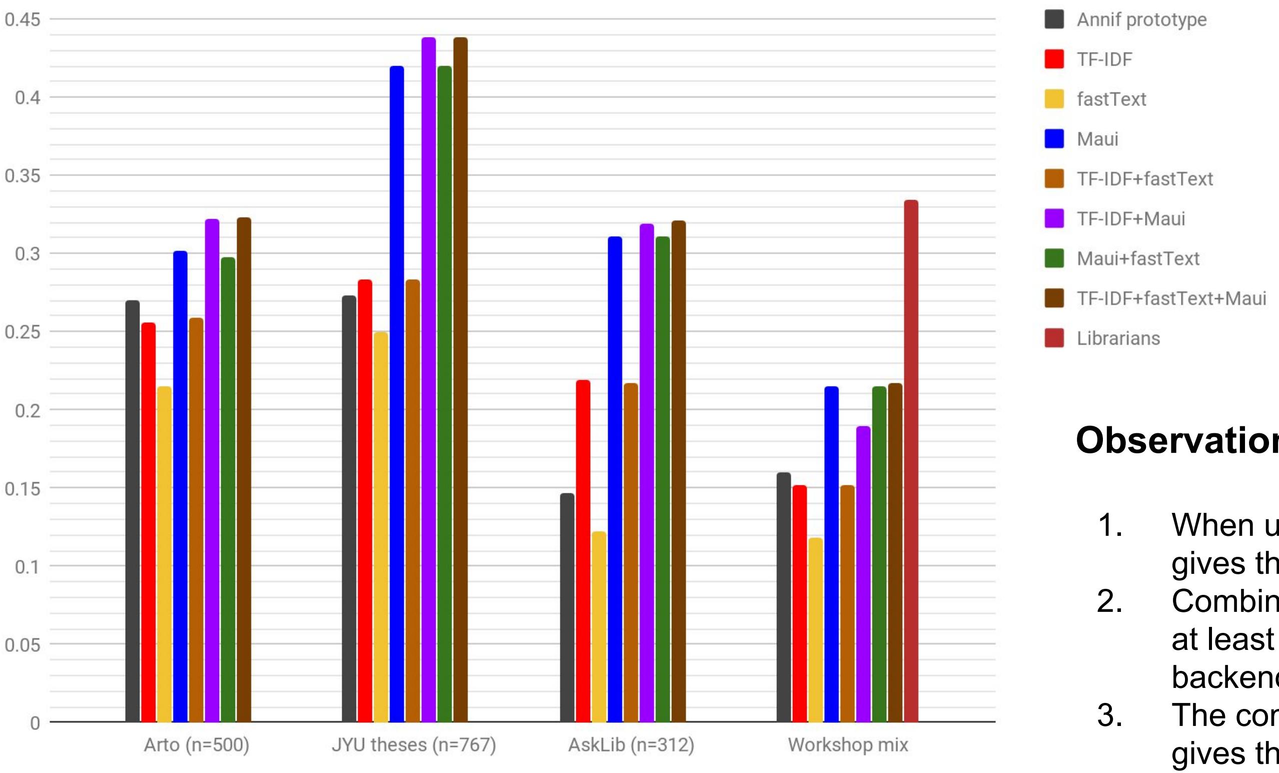
- Question/Answer pairs from an Ask a Librarian service (n=3150)
  Short, informal questions and answers about many different topics.

Available on GitHub: https://github.com/NatLibFi/Annif-corpora

(for the first two corpora, only links to PDFs are provided for copyright reasons)

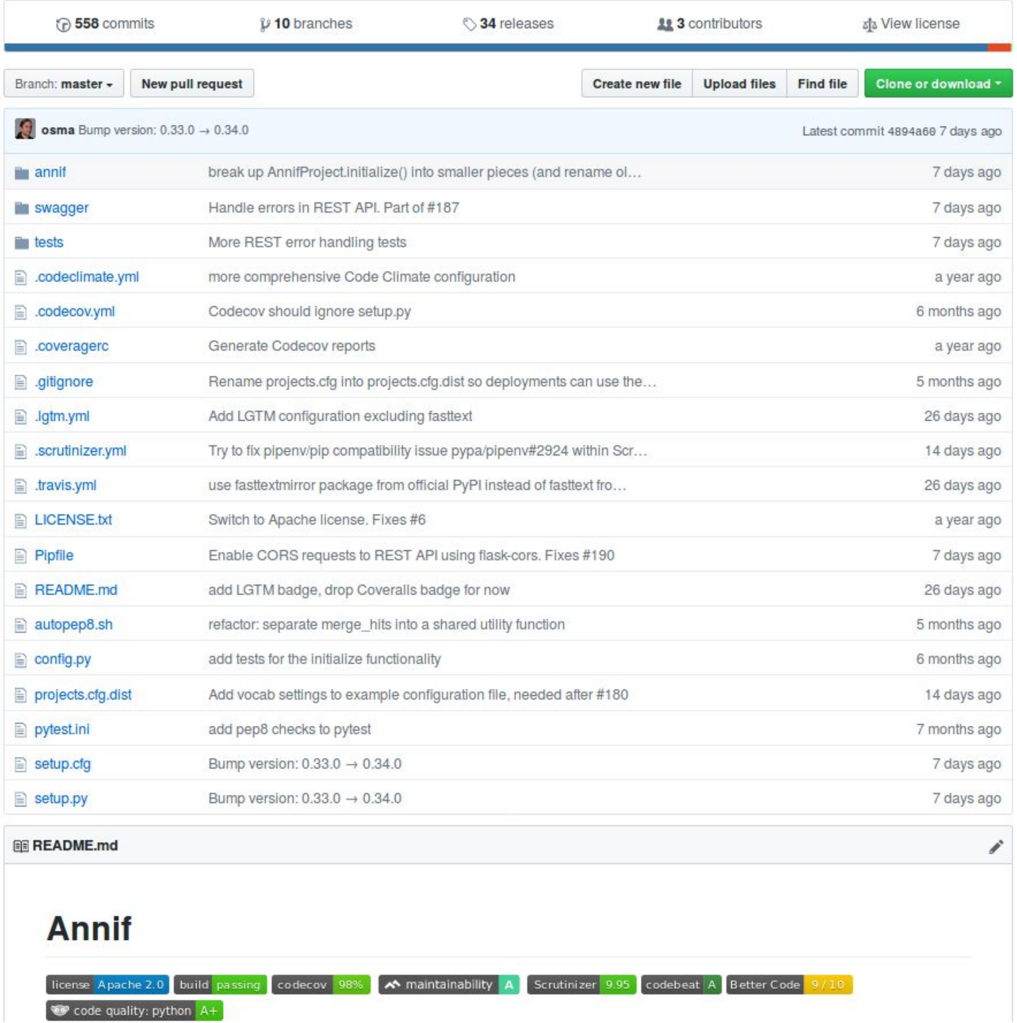
Evaluation of different backends

F-measure similarity scores against a gold standard



Observations:

1. When using just one backend, Maui often gives the best results
2. Combinations (ensembles) usually give at least as good results as single backends
3. The combination of all three backends gives the best results



osma Bump version: 0.33.0 → 0.34.0 Latest commit 4894a60 7 days ago

annif	break up AnnifProject.initialize() into smaller pieces (and rename ol…	7 days ago
swagger	Handle errors in REST API. Part of #187	7 days ago
tests	More REST error handling tests	7 days ago
.codeclimate.yml	more comprehensive Code Climate configuration	a year ago
.codecov.yml	Codecov should ignore setup.py	6 months ago
.coveragerc	Generate Codecov reports	a year ago
.gitignore	Rename projects.cfg into projects.cfg.dist so deployments can use the…	5 months ago
.lgtm.yml	Add LGTM configuration excluding fasttext	26 days ago
.scrutinizer.yml	Try to fix pipenv/pip compatibility issue pypa/pipenv#2924 within Scr…	14 days ago
.travis.yml	use fasttextmirror package from official PyPI instead of fasttext fro…	26 days ago
LICENSE.txt	Switch to Apache license. Fixes #6	a year ago
Pipfile	Enable CORS requests to REST API using flask-cors. Fixes #190	7 days ago
README.md	add LGTM badge, drop Coveralls badge for now	26 days ago
autopep8.sh	refactor: separate merge_hits into a shared utility function	5 months ago
config.py	add tests for the initialize functionality	6 months ago
projects.cfg.dist	Add vocab settings to example configuration file, needed after #180	14 days ago
pytest.ini	add pep8 checks to pytest	7 months ago
setup.cfg	Bump version: 0.33.0 → 0.34.0	7 days ago
setup.py	Bump version: 0.33.0 → 0.34.0	7 days ago

README.md

Annif

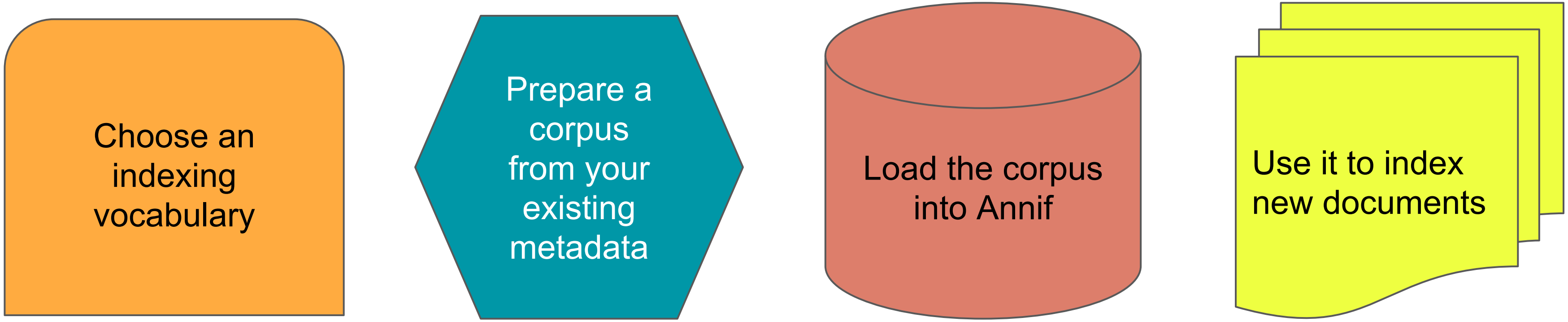
Annif on GitHub

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

Apply Annif on your own data!



Choose an indexing vocabulary

Prepare a corpus from your existing metadata

Load the corpus into Annif

Use it to index new documents

Lessons learned (so far)

1. Good quality training data is key for training and evaluation

   Don’t expect good results if you don’t have the data it takes

2. Gold standard subjects are useful, but human evaluation is necessary

   Subject indexing is inherently subjective; comparing to a single gold standard can be misleading

3. All algorithms have strong and weak points

   Combinations work better than any algorithm by itself

4. Surprising amount of interest also from non-library organizations

   Archives, media organizations, book distributors … automation is better done together!

Thank you!

Questions?

osma.suominen@helsinki.fi - @OsmaSuominen

Website: http://annif.org
Code: https://github.com/NatLibFi/Annif
Test corpora: https://github.com/NatLibFi/Annif-corpora

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