Annif and Finto AI:
Developing and implementing automated subject indexing

Osma Suominen, Mona Lehtinen, Juho Inkinen
International Conference on Bibliographic Control in the Digital Ecosystem
10 February 2021
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 40,000+ subjects (including places)
YSA
YSO
Allärs
KOKO
black box
Machine learning using existing metadata

The collections of Finnish archives, libraries and museums at your fingertips.

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 software
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
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."

Lexical approaches need comparatively little training data.

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

learn which **subjects** are correlated with which **words**
in documents, based on training data

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

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)

3. Community building
The conference aims to explore the new boundaries of Universal bibliographic control. Bibliographic control is radically changing because the bibliographic universe is radically changing: resources, agents, technologies, standard, and practices. As a “non-commercial public space” (IFLA Global vision) - not only in a literal sense - libraries still play a fundamental role also in the digital ecosystem.

Among the topics that will be addressed in the Conference:
- the new bibliographic universe;
- library cooperation networks;
- the legal deposit;
- national bibliographies;
- bibliographic agencies;
- the new control tools and standards (IFLA LRM, RDA, BIBFRAME);
- authority control and new alliances: Wikidata, Wikibase, Identifiers;
- new ways of indexing documents (artificial intelligence, machine learning, text-mining);
- the role of thesauri and ontologies in the digital ecosystem;
- changes in the coverage area of bibliographic control by libraries
- bibliographic control and search engines.
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
Hands-on **Annif tutorial** for those who want to use Annif on their own

**Videos and exercises freely available on YouTube & GitHub!**
4. Annif deployments
JYX repository, University of Jyväskylä
Students upload their Master’s and doctoral theses, Annif suggests subjects*

Keywords

- 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

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

Launched in May 2020

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 collections of machines) that interact and associate with the human mind, such as through speech or writing.

As machines become increasingly capable, the more complex tasks traditionally performed by humans are increasingly taken over by AI, a development that some observers view with alarm. In 2006, The New York Times reported that "AI is whatever we don't yet know how to program.[2]" Artificial intelligence has been an active research area for more than six decades, and many academic fields and commercial applications of AI exist, including search, speech recognition, medical diagnosis, computer vision, game playing, natural language processing, robotics, and more.

The field of AI has been compared to the Turing test. The Turing Theorem says "AI is whatever we don't yet know how to program,"[3] meaning that every time we think we have created a machine that can think like a human, we are actually defining the machine to think like us.

Modern machine capabilities generally classified as AI include successfully understanding human speech,[4] competing at the highest level in strategic game systems (such as chess and Go),[5] autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

ai.finto.fi
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

Publishers

information about new titles

Kirjavälitys

? correction and curation

Melinda inc. Fennica

Libraries

Bookshops and online stores

descriptive text

subject suggestions

fintoai annif
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...

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

What is this?
What should I do here?

Maybe it’s better to leave these alone...
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)


Endorsements

"C’est cette spécificité qu’Annif fournit. Le logiciel a été conçu comme une surcouche à Tensorflow (et autres fonctions incluses) précisément adaptée à la fonction d’indexation. Cela ne garantit pas en soi l’efficacité du processus, mais ça facilite grandement le travail pour pouvoir le tester."
-- Etienne Cavalié (Lully), BNF (blog post)

"Through a proof-of-concept, we have shown that using Annif would be a promising approach to improve the searchability of this collection through automatic categorisation."
-- Sara Veldhoen, KB.nl (blog post)

"It cannot be worse than the hack we had before!"
-- Antonin Delpeuch, Dissem.in (forum post)
Thank you!

Juho Inkinen
Mona Lehtinen
Osma Suominen

annif.org


These slides: https://tinyurl.com/annif-bc2021