NATIONAL LIBRARY OF FINLAND

Annif and automated subject cataloguing

annif

developed since 2017

- General purpose open source **tool** for automated subject indexing and classification
- Multilingual, supports many vocabularies
- Code on GitHub, website with test form and API
- Global development and user community; user forum **annif-users** on Google Groups

annif.org

1640 NATIONAL LIBRARY OF FINLAND

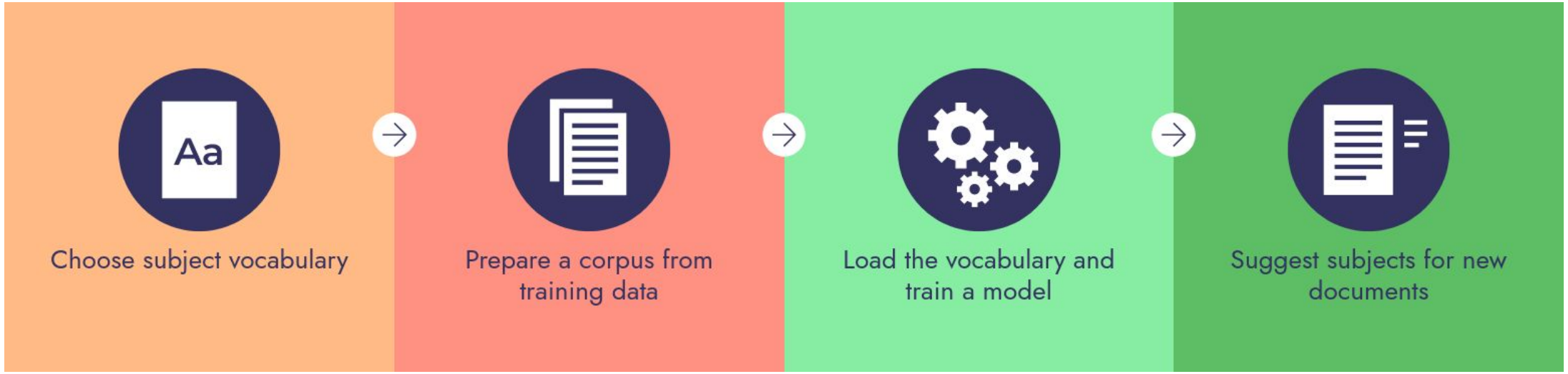
fintoai

launched in 2020

- Automated subject indexing **service** for production use, based on Annif. Web user interface and API service
- Supports indexing with the General Finnish Ontology YSO & PLC – Finnish Public Libraries Classification System (in Fin, Swe & Eng) as well as KAUNO – ontology for fiction (in Finnish)
- Intended to support subject cataloguers in Finland regardless of institution (GLAMs, public administration); sister project to the Finto vocabulary service

ai.finto.fi

What would it take to use Annif?



Corpora

- https://github.com/NatLibFi/Annif-corpora
- https://github.com/NatLibFi/Annif/wiki/Corpus-formats
- https://github.com/NatLibFi/Annif/wiki/Achieving-good-results

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

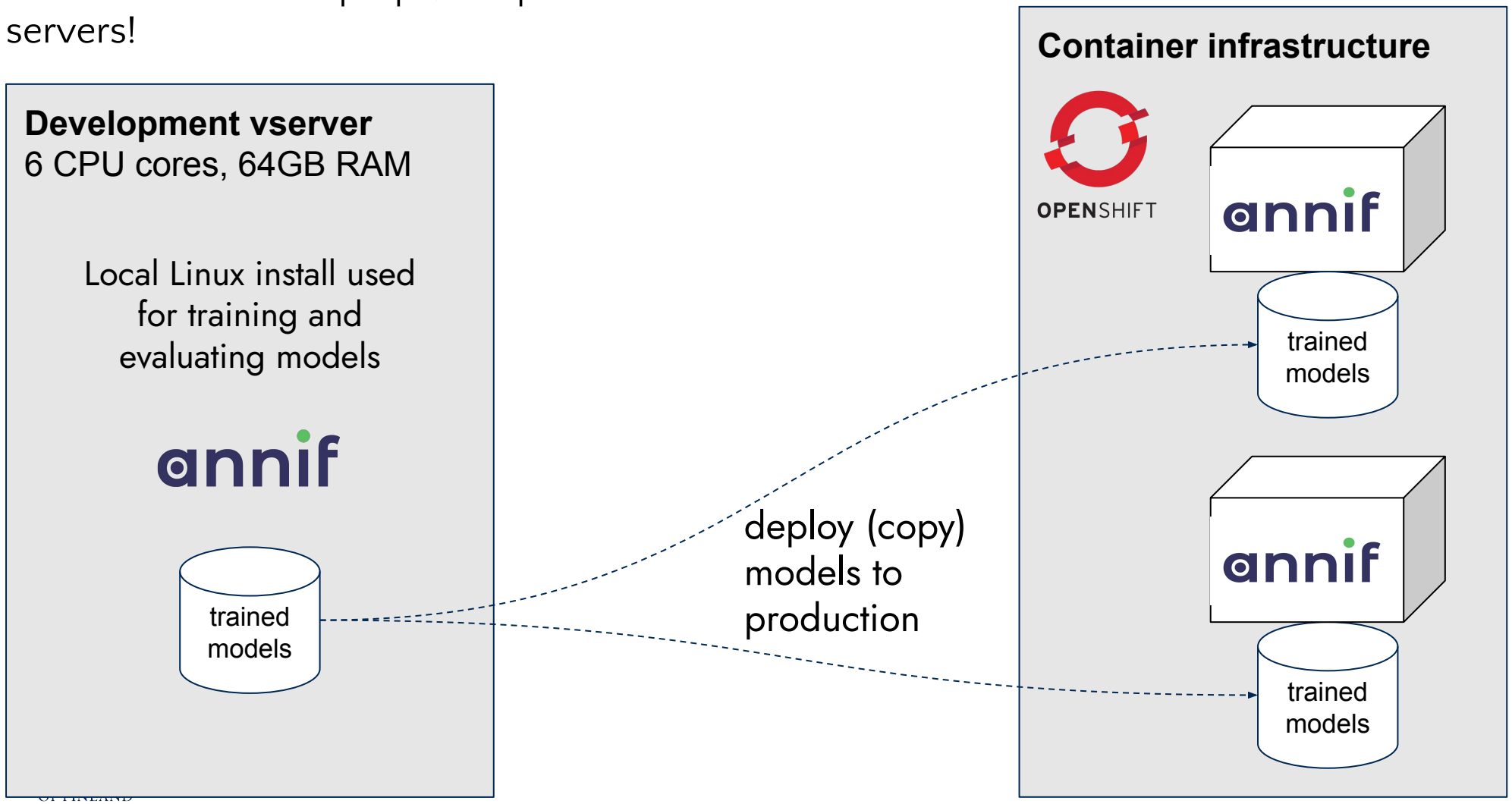
1. Evaluating indexing quality directly through **assessment by an evaluator** or by **comparison with a gold standard**.
2. Evaluating indexing quality directly **in the context of an indexing workflow**.
3. Evaluating indexing quality indirectly through retrieval performance.

The different evaluation approaches are complementary.
Not a good idea to look at just a single measure.

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.

Technical infrastructure for production use

You can start with laptops, but production use needs servers!



fintoai

api.annif.org

Annif tutorial

For hands-on details of working with Annif, see the Annif tutorial

Most recent Annif tutorial / workshop: Open Repositories 2024 conference in Göteborg, Sweden, 3–6 June 2024

Materials (videos and exercises) are available for self-study on GitHub and YouTube.

Partners and users

Collecte

yle

KB nationale bibliotheek

CSC

KirjaVälitys

KANSALLISARKISTO

Dissemin

University of Jyväskylä | JYX Digital Repository

National Library of Sweden

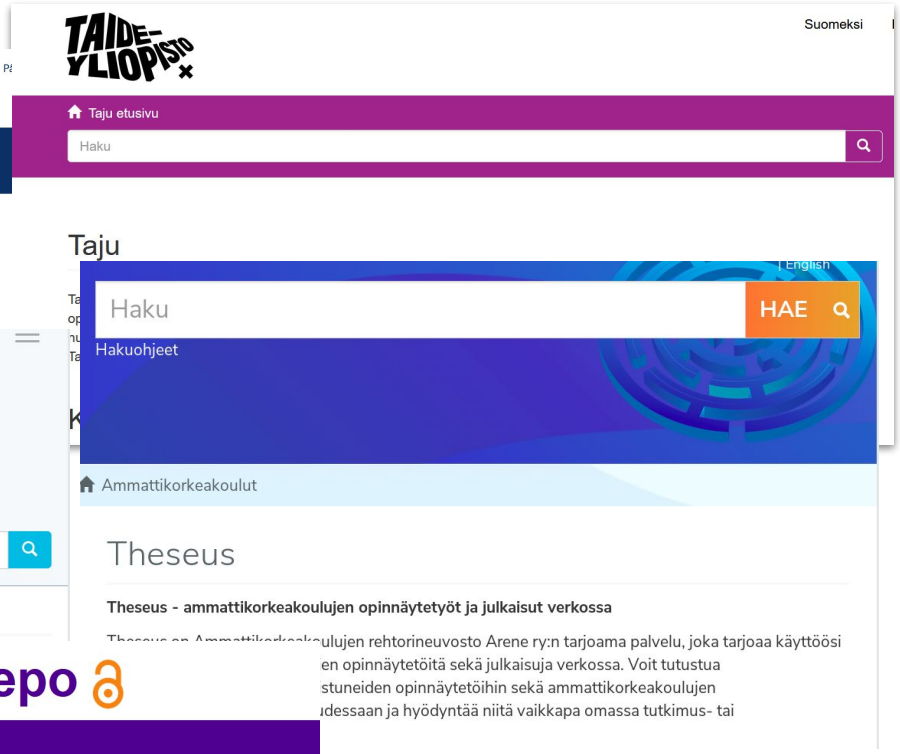
ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

KANSALLINEN AUDIOVISUAALINEN INSTITUUTTI
NATIONELLA AUDIOVISUELLA INSTITUTET
NATIONAL AUDIOVISUAL INSTITUTE

DEUTSCHE NATIONAL BIBLIOTHEK

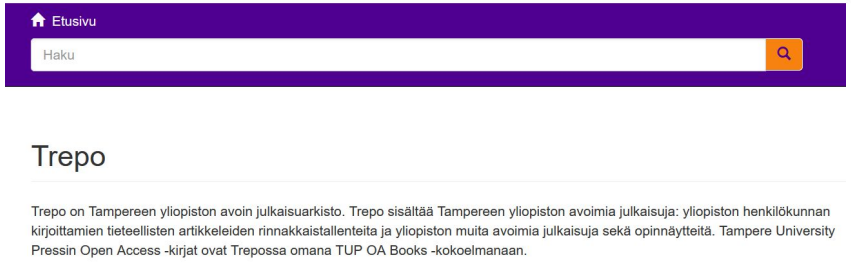
KIRJASTOT.fi

Usage at NatLibFi



Legal deposits

Trepo



16 40
KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

DSpace-based institutional repositories: Osuva, Trepo, Theseus, Taju, Lauda, Kaisu (Demo of use)

Current & future developments

- Automated language detection for Finto AI service
- New vocabularies
- The role of AI in subject indexing and in the library field
- User survey:
  - all in all good grades
- Automatic extraction of bibliograph. metadata
  https://github.com/NatLibFi/FinGreyLit/tree/main/experiments

Extracting biblographic metadata

PDFs

system

```
title: Novel methods for earth fault passage indication in non-effectively grounded electricity distribution networks

contributor/faculty: fi=School of Technology and nnovations|en=School of Technology and Innovations|

contributor/author: Farughian, Amir

contributor/organization: fi=Vaasan yliopisto|en=University of Vaasa|

publisher: Vaasan yliopisto

date/issued: 2022

relation/issn: 2323-9123

relation/issn: 0355-2667

relation/isbn: 978-952-395-054-2

relation/ispartofseries: Acta Wasaensia

relation/numberinseries: 500
```

16 40
KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

Meteor+LLM results?

field	Meteor	Qwen2-0.5B	Mistral-7B
language	97%	98%	99%
title	40%	73%	87%
alt_title	76%	73%	78%
creator	65%	72%	88%
publisher	8%	69%	83%
year	72%	87%	89%
e-isbn	83%	90%	96%
e-issn	86%	95%	95%
p-isbn	72%	89%	93%
p-issn	83%	92%	94%
doi	91%	97%	98%
type_coar	0%	85%	90%
	64%	**85%**	**90%**

Material used: FinGreyLit Corpus

- 800 PDFs + metadata
- 9 DSpace repositories
- 3 langs (fin, swe, eng + sme)
- theses, e-books, articles, reports etc.

Meteor: original Meteor tool

Qwen2-0.5B: smaller LLM, 500 perams CPU

Mistral-7B: larger LLM, 7 billion params. GPU

See also:
https://forum.swib.org/t/automating-metadata-extraction-and-cataloguing-experiences-from-the-national-libraries-of-norway-and-finland/121

Further reading (& use cases)

- An article about Annif and Finto AI has been published in 2022 in the peer-reviewed Open Access journal JLIS.it.

Further reading: preprocessing text

Annif supports many language-specific Analyzers.

- **simple** – simple but stupid, no stemming or lemmatization
- **simplemma** – simple lemmatization library, supports Estonian

These and other approaches need to be evaluated.

See Code4Lib Journal article about different lemmatizers:
Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing

NATIONAL LIBRARY OF FINLAND

Thank you!

Mona Lehtinen, Osma Suominen, Juho Inkinen

[firstname.lastname]@helsinki.fi