Automated subject indexing with Annif and Finto AI

Putting DIY automated subject indexing into production

Osma Suominen, Mona Lehtinen, Juho Inkinen

DCMI Virtual AI panel discussion
13 October 2021

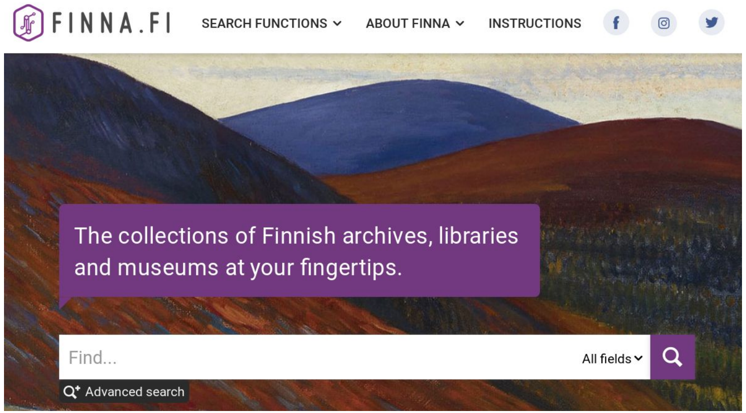
16 40
NATIONAL LIBRARY
OF FINLAND

Outline

1. Development of Annif
2. Quality of automated subject indexing
3. Community building
4. Annif deployments
5. Lessons learned

1. Development of Annif

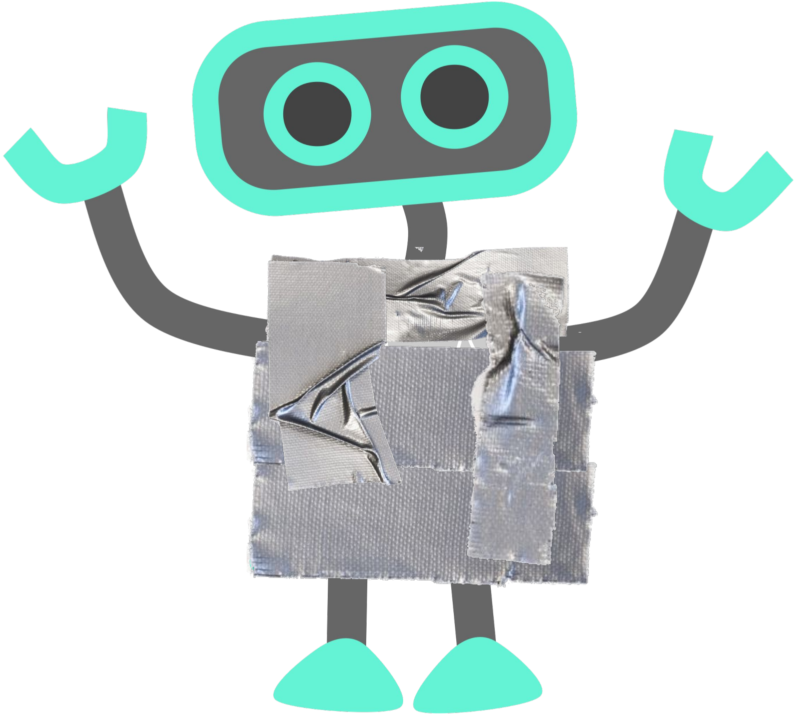
Machine learning using existing metadata



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

annif

Early prototype (2017) got people excited



Goals for Annif implementation (2018 →)

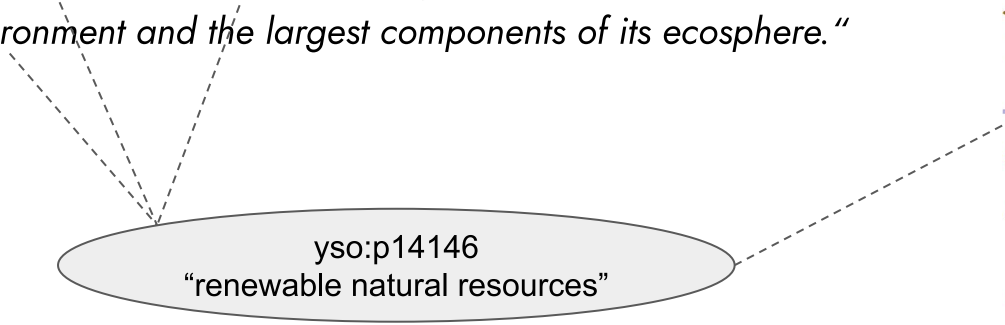
1. multilingual
2. independent of indexing vocabulary
3. support different subject indexing algorithms
4. CLI, Web user interface and REST API
5. community-oriented open source software

Lexical vs. associative algorithms for subject indexing

lexical approaches (e.g.: Maui, MLLM, STWFSA)

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.: fastText, Omikuji, SVC)

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

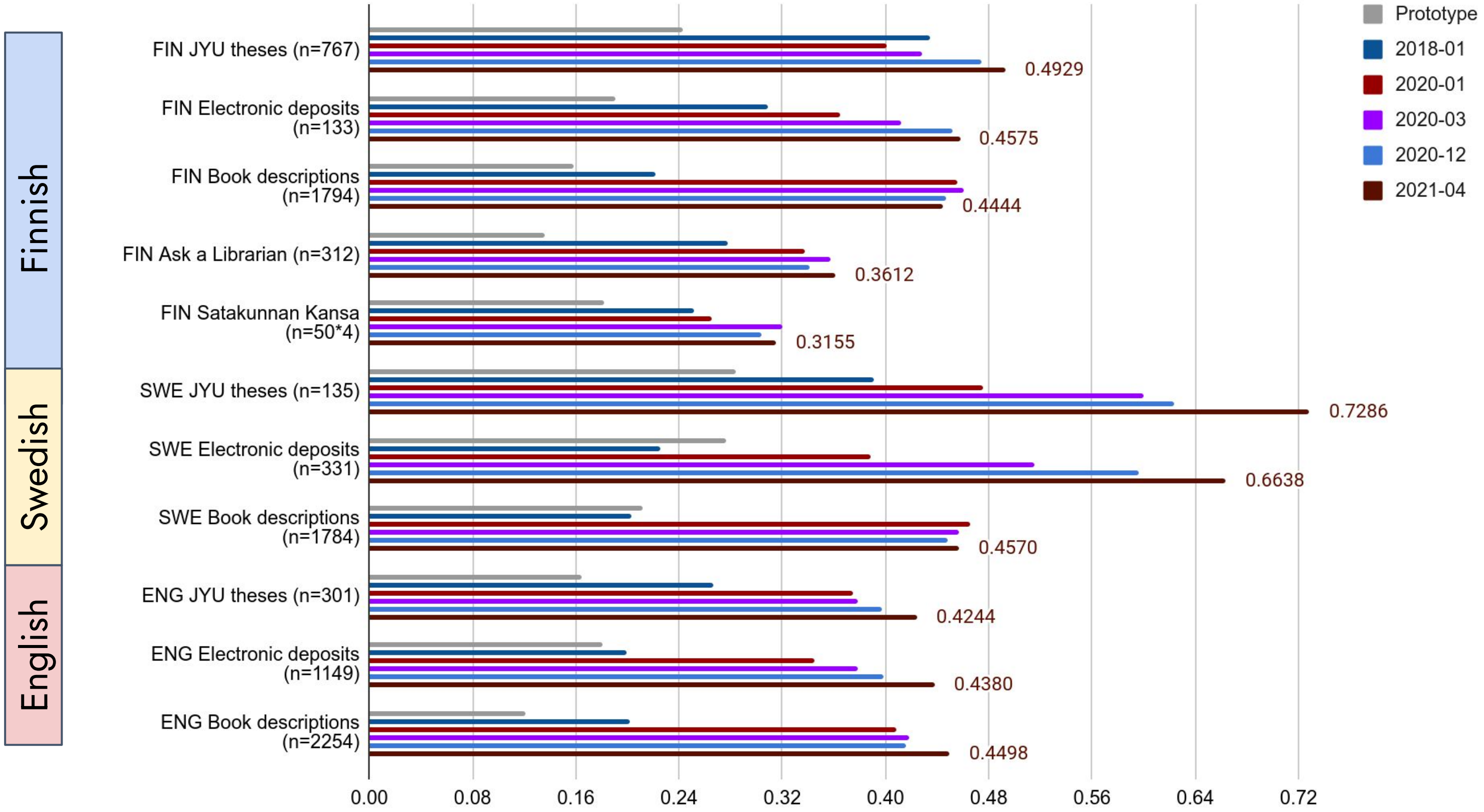
renew
energi
resourc
environment
sustain
develop
sourc
technolog
product
natur
photovolta
report
manag
implement
northern
assess
system
polici
biomass
solar
econom
finland
model
pollut
practic
perspect
design
analysi
region
global
summari
local
power
forest
new
potenti
action
plan
thermal
architectur
commun
chang
hydrogen
opportun
research
challeng
approach
programm
carbon
chemic
futur
electr
impact
biotechnolog
agricultur
fuel
project
wood
book
growth
proceed
petroleum
control
measur
rural
sector
africa
plant
water
environ
india
cell
fibr
econom
gener
altern
union
option
beyond
russia
ecolog
statist
handbook
materi
execut
base
alberta
overlap
solut
oil
stock
case
affair
capit
compani
toward
end
eu

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

2. Quality of automated subject indexing

Comparison to “gold standard”

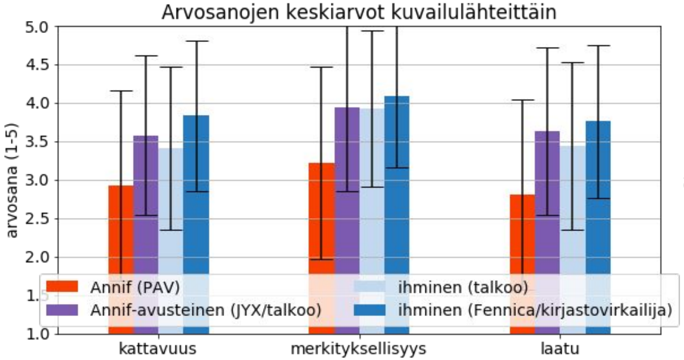
F1@5 scores for different test corpora and Annif API/model versions



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).



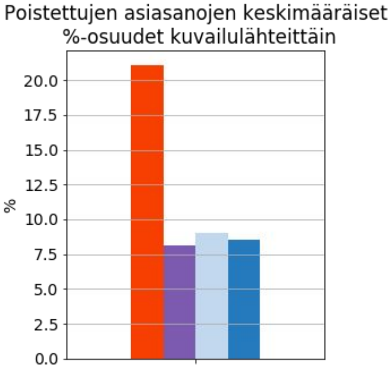
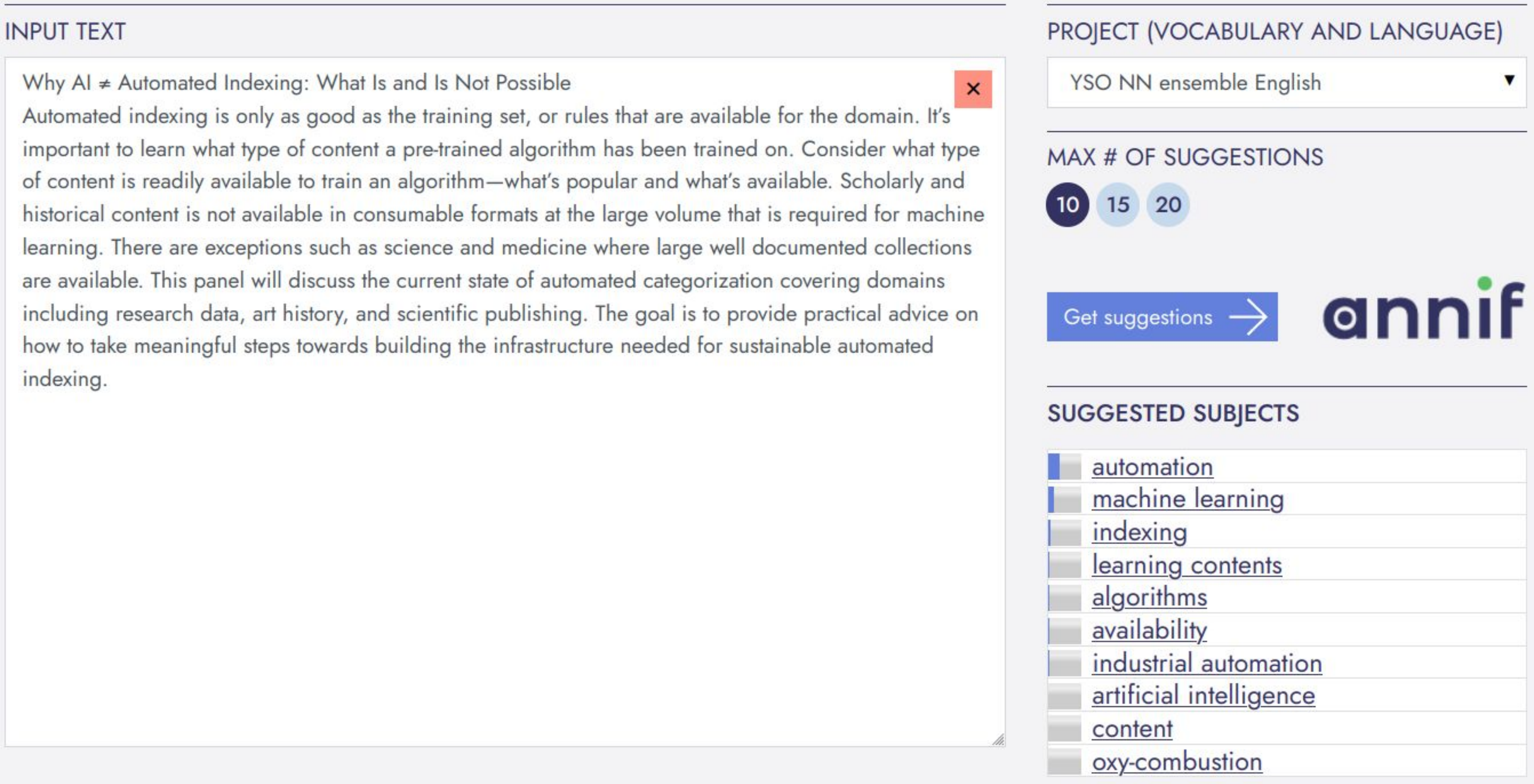


Photo: Mikko Lappalainen.

Lehtinen M., Inkinen J. & Suominen O. (2019). Aaveita koneessa: Automaattisen sisällönkuvailun arviointia Kirjastoverkkopäivillä 2019. *Tietolinja, 2019(2)*. http://urn.fi/URN:NBN:fi-fe2019120445612

3. Community building

Web site with form for testing at annif.org



INPUT TEXT

Why AI ≠ Automated Indexing: What Is and Is Not Possible

Automated indexing is only as good as the training set, or rules that are available for the domain. It's important to learn what type of content a pre-trained algorithm has been trained on. Consider what type of content is readily available to train an algorithm—what's popular and what's available. Scholarly and historical content is not available in consumable formats at the large volume that is required for machine learning. There are exceptions such as science and medicine where large well documented collections are available. This panel will discuss the current state of automated categorization covering domains including research data, art history, and scientific publishing. The goal is to provide practical advice on how to take meaningful steps towards building the infrastructure needed for sustainable automated indexing.

PROJECT (VOCABULARY AND LANGUAGE)

YSO NN ensemble English ▾

MAX # OF SUGGESTIONS

● 10 ○ 15 ○ 20

SUGGESTED SUBJECTS

- automation
- machine learning
- indexing
- learning contents
- algorithms
- availability
- industrial automation
- artificial intelligence
- content
- oxy-combustion

Hands-on Annif tutorial

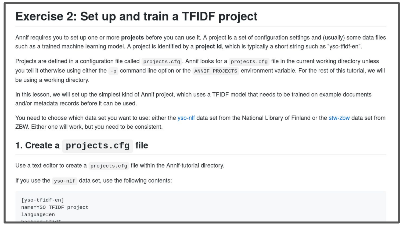
for those who want to use Annif on their own

SWIB19
Semantic Web in Libraries

DCMI Virtual, 2020
September 14th-25th, 2020

SWIB20
Semantic Web in Libraries

annif tutorial
Data se
tutorial
1640 THE NATIONAL LIBRARY OF FINLAND
ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics



Exercise 2: Set up and train a TFIDF project

Annif requires you to set up one or more **projects** before you can use it. A project is a set of configuration settings and (usually) some data files such as a trained machine learning model. A project is identified by a **project id**, which is typically a short string such as "yso-tfidf-en".

Projects are defined in a configuration file called `projects.cfg`. Annif looks for a `projects.cfg` file in the current working directory unless you tell it otherwise using either the `-p` command line option or the `ANNIF_PROJECTS` environment variable. For the rest of this tutorial, we will be using a working directory.

In this lesson, we will set up the simplest kind of Annif project, which uses a TFIDF model that needs to be trained on example documents and/or metadata records before it can be used.

You need to choose which data set you want to use: either the yso-nlf data set from the National Library of Finland or the stw-zbw data set from ZBW. Either one will work, but you need to be consistent.

1. Create a `projects.cfg` file

Use a text editor to create a `projects.cfg` file within the Annif-tutorial directory.

If you use the `yso-nlf` data set, use the following contents:

```
[yso-tfidf-en]
name=YSO TFIDF project
language=en
backend=tfidf
```

Videos and exercises freely available on YouTube & GitHub!

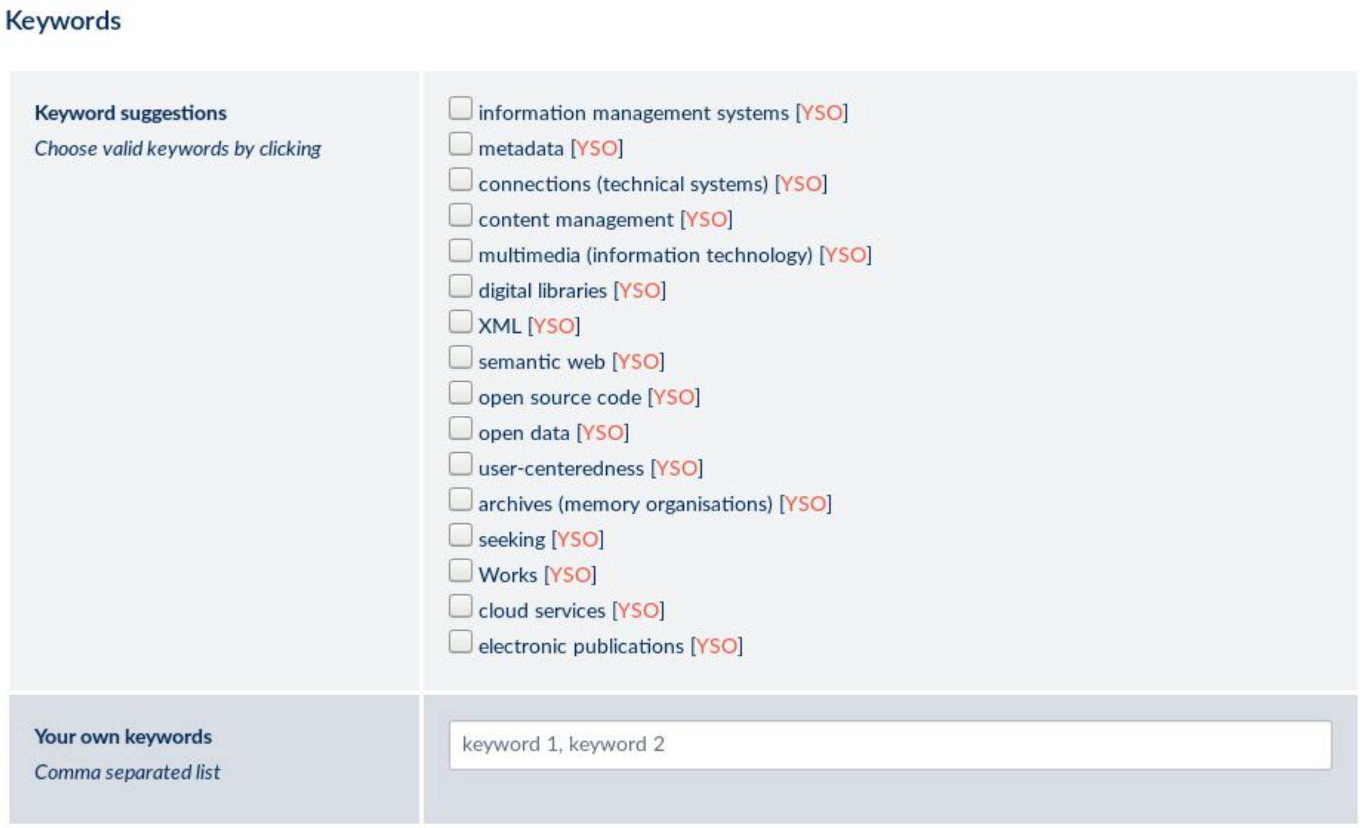
1640
NATIONAL LIBRARY
OF FINLAND

ZBW
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

4. Annif deployments

JYX repository, University of Jyväskylä

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



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

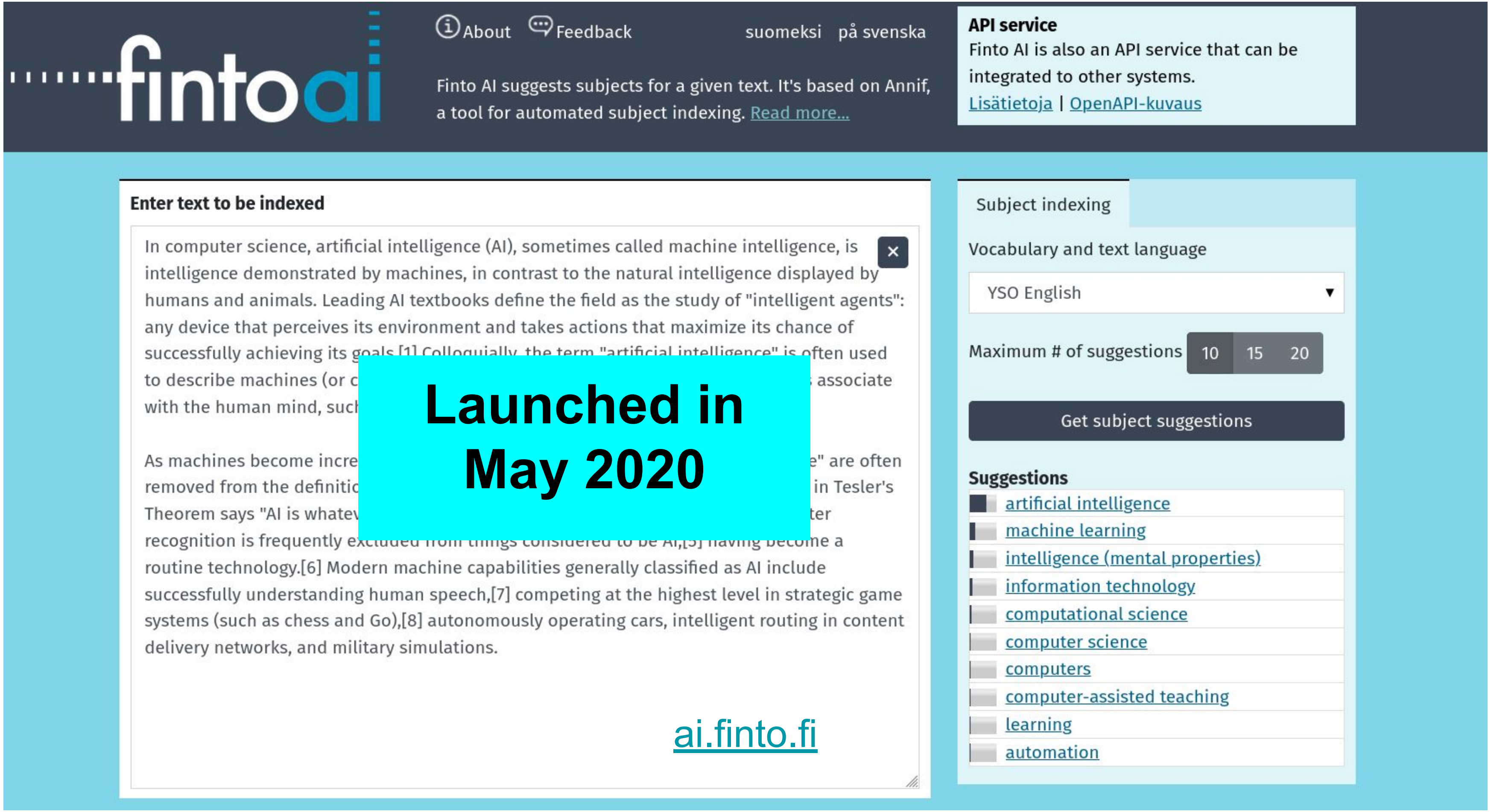
Comma separated list

keyword 1, keyword 2

Implemented using DSpace & GLAMpipe by Ari Häyrinen

*from YSO = General Finnish Ontology

Finto AI - automated subject indexing tool and API service



Finto AI suggests subjects for a given text. It's based on Annif, a tool for automated subject indexing. Read more...

API service

Finto AI is also an API service that can be integrated to other systems.

Lisätietoja | OpenAPI-kuvaus

Enter text to be indexed

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 c … associate with the human mind, such …

As machines become incre … e" are often removed from the definitio … in Tesler's Theorem says "AI is whatev … ter recognition is frequently excluded from things considered to be AI,[5] having become a routine technology.[6] Modern machine capabilities generally classified as AI include successfully understanding human speech,[7] competing at the highest level in strategic game systems (such as chess and Go),[8] autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

Vocabulary and text language: YSO English ▼

Maximum # of suggestions: 10 | 15 | 20

Suggestions

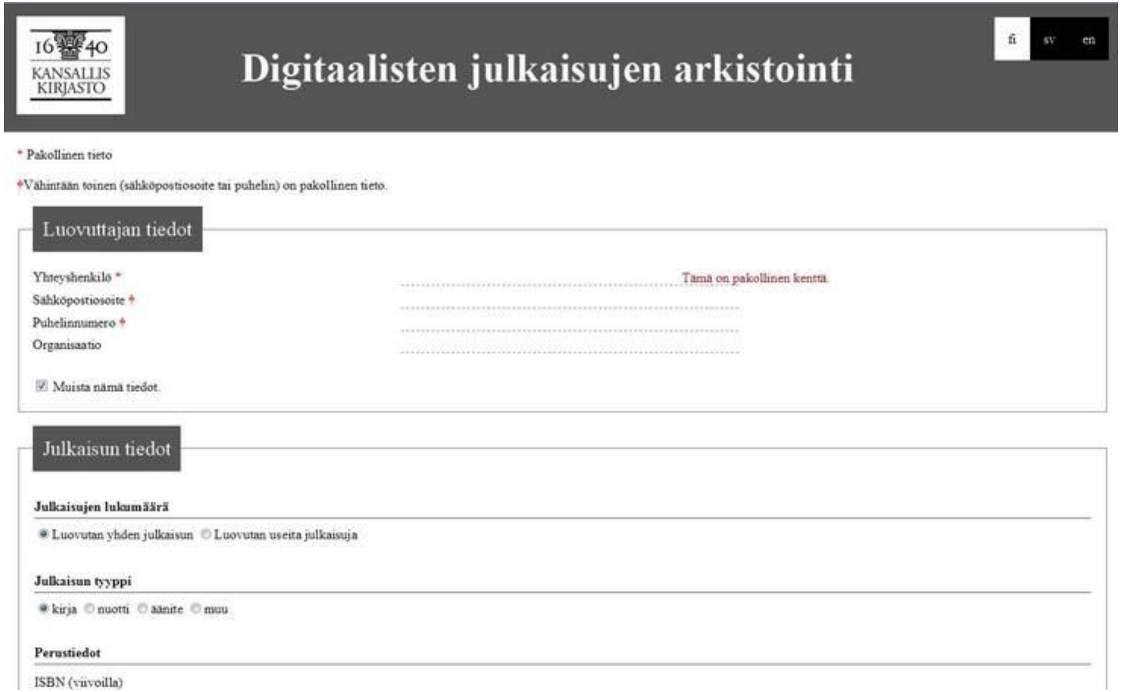
- artificial intelligence
- machine learning
- intelligence (mental properties)
- information technology
- computational science
- computer science
- computers
- computer-assisted teaching
- learning
- automation

Launched in May 2020

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.



Digitaalisten julkaisujen arkistointi

* Pakollinen tieto

*Vähintään toinen (sähköpostiosoite tai puhelin) on pakollinen tieto.

Luovuttajan tiedot

Yhteyshenkilö * .... Tämä on pakollinen kenttä.

Sähköpostiosoite * ....

Puhelinnumero * ....

Organisaatio ....

☑ Muista nämä tiedot.

Julkaisun tiedot

Julkaisujen lukumäärä

● Luovutan yhden julkaisun ○ Luovutan useita julkaisuja

Julkaisun tyyppi

● kirja ○ nuotti ○ äänite ○ muu

Perustiedot

ISBN (viivoilla)

Implementation: Erik Lindgren, Mikko Merioksa, Satu Niininen

5. Lessons learned

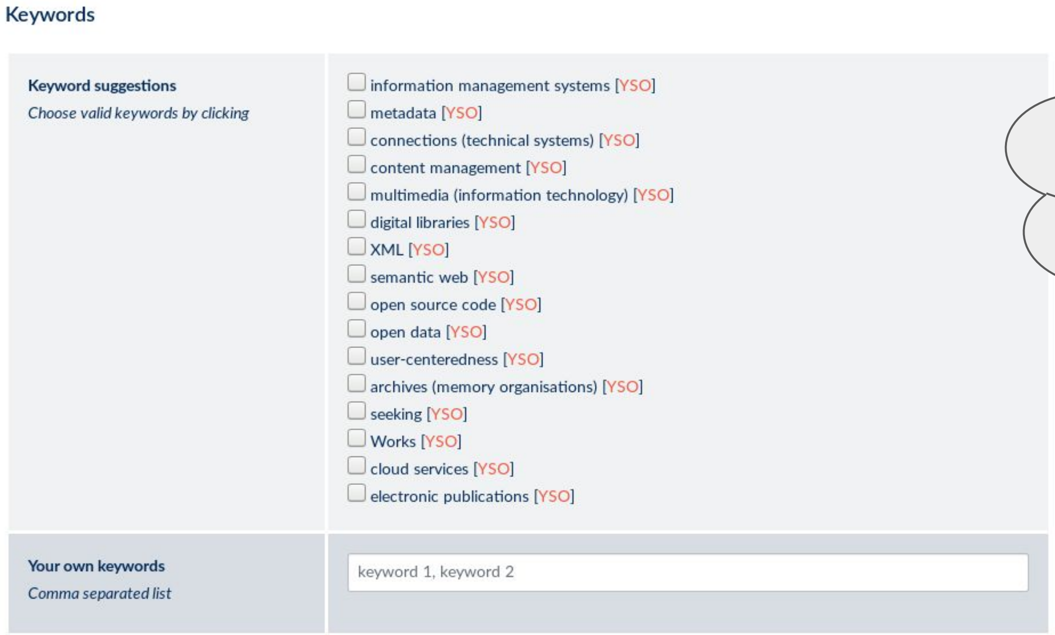
Algorithms may be used **alone**, or in combinations, **ensembles**

Ensembles are nearly always better than individual algorithms

Start by experimentation, move slowly towards production

image credit: @kettudolls (IG)

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…

Thank you!

Juho Inkinen

Mona Lehtinen

Osma Suominen

annif.org

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