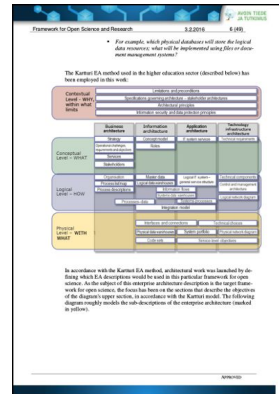
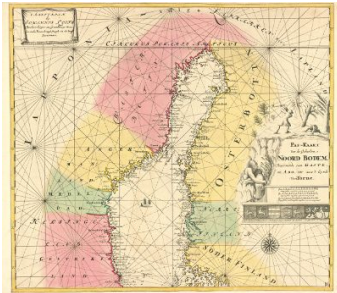
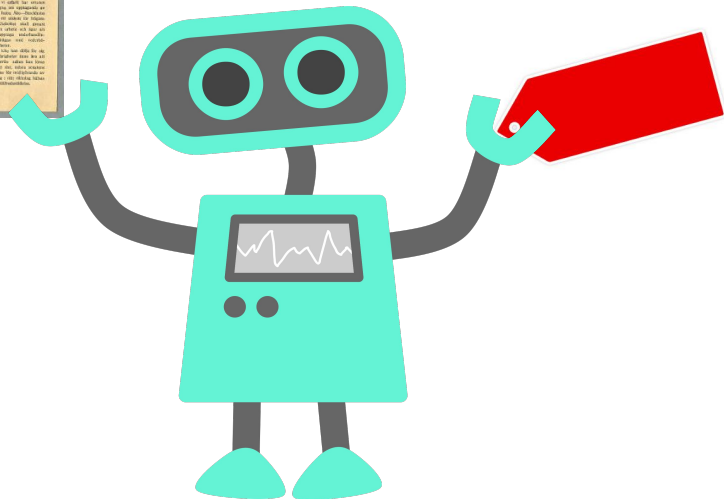


Annif - työkalu automaattiseen indeksointiin ja luokitteluun

Osma Suominen

BMF syysseminaari 22.11.2019





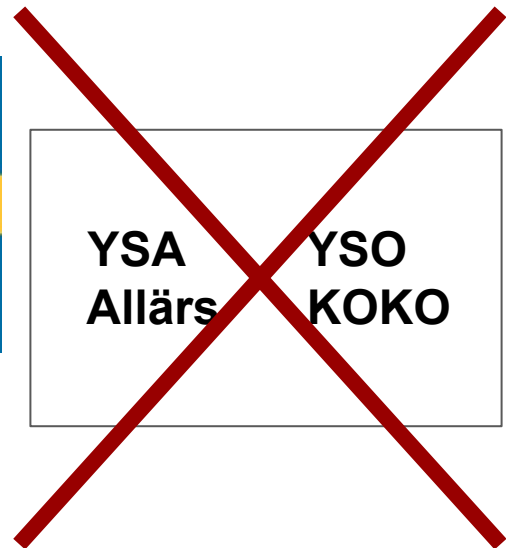
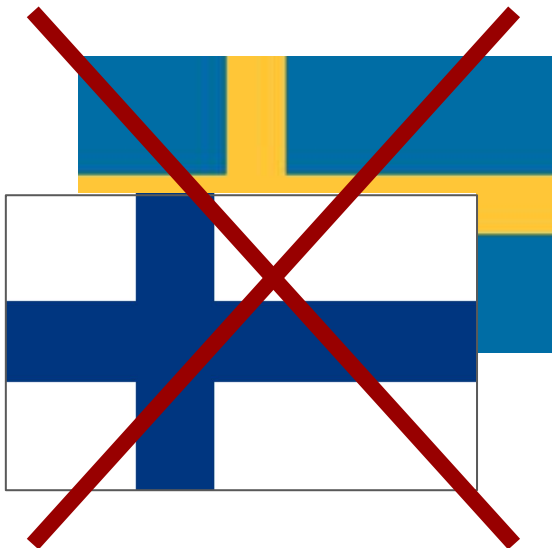


OPEN
CALAIS



THOMSON REUTERS





€£\$

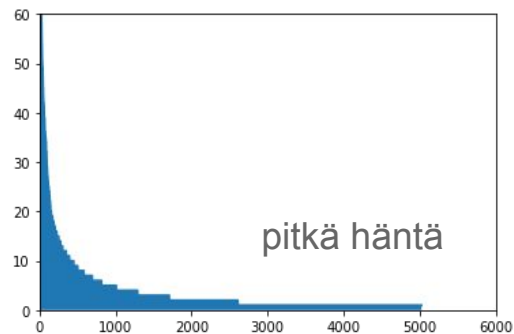
Sisällönkuvailu on vaikeaa

ihmisille:

- **Subjektiivisuus:** kun kaksi eri ihmistä kuvailee saman dokumentin, vain $\sim\frac{1}{3}$ aiheista on samoja
- **Paljon käsitteitä:** kymmeniä tuhansia mahdollisia aiheita joista valita
- **Sanasto elää:** uusia käsitteitä lisätään, vanhoja nimetään ja määritellään uudelleen

tekoälylle:

- **Pitkän hännän jakauma:** useimpien aiheiden käytöstä vain vähän esimerkkejä
- **Paljon käsitteitä:** vaatii monimutkaisia ja laskennallisesti raskaita malleja
- **Vaikea arvioida:** hyvien ehdotusten erottelu huonoista vaatii ihmistyötä
- **Sanasto elää:** malleja täytyy kouluttaa uudelleen



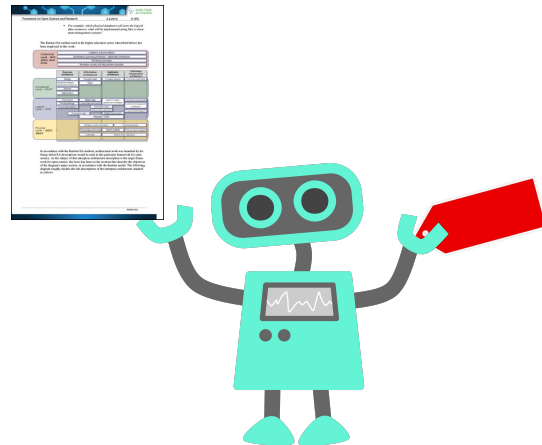
Annif

Tuottaa sisällönkuvailua (asiasanoitus ja luokitus)
tekstin perusteella

Kansalliskirjastossa kehitetty työkalu, avointa koodia

Soveltuu parhaiten asiatekstille, mutta testataan myös muilla aineistoilla

Perustuu kieliteknologiaan ja koneoppimiseen



Suomen arkistojen, kirjastojen ja museoiden
aarteet samalla haullla.

Hae...

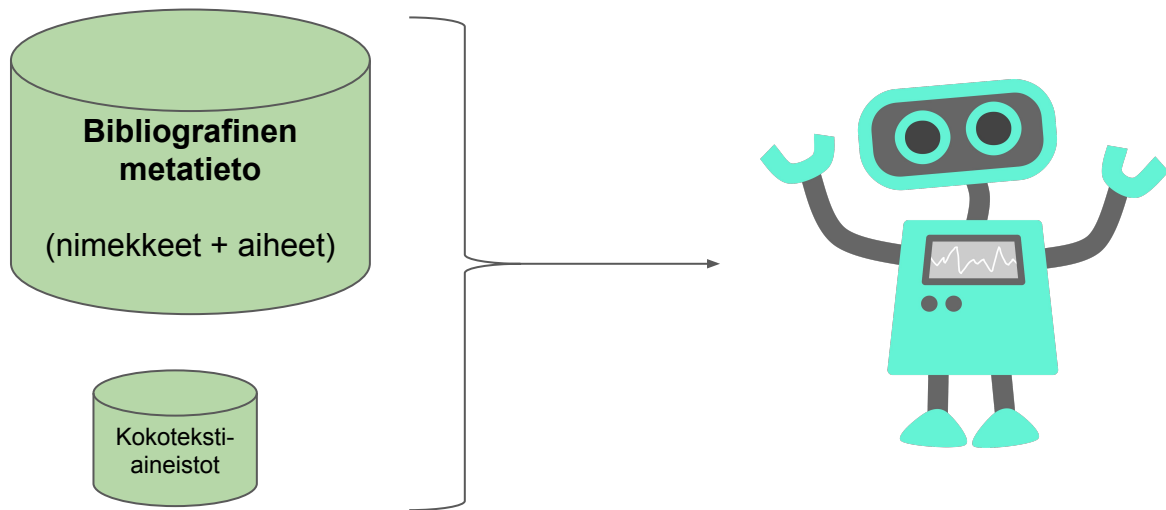
Kaikki osumat ▾



🔍 Tarkennettu haku

Paljon muistiorganisaatioiden metatietoa, mm. **15M tietuetta** [Finnassa](#)

Annif oppii aineistosta, miten tunnistaa aiheita



Lomake testausta varten: annif.org

Try Annif!

Text to analyze:

Miksi tuntuu kuin muut tuijottaisivat vihaisesti tai katsovat ohi? Tuore väitös selittää, miten tulkitsemme "väärin" muita ihmisiä

Yksin jääminen muuntaa tuoreen väitöksen mukaan reaktioita toisten näyttämiin sosiaalisiin viesteihin.

Miksi toisten katseet tuntuvat niin pahalta? Tuntuuko, että olet ulkopuolinen tai kuin sinua ei olisi olemassakaan?

Psykologian maisteri Aleksi Syrjämäen väitöskirjassa havaittiin, että yksin jääminen muuntaa reaktioita toisten näyttämiin sosiaalisiin viesteihin.

Väitöskirja osoitti, että yksin jääneestä saattaa tuntua siltä, etteivät muut edes katso häntä kohti. Toisaalta vaikka joku katsoisi kohti, tämäkään ei välttämättä piristä yksinäistä.

Reaktio voi kertoa halusta suojautua

Project (vocabulary and language):

YSO ensemble Finnish

Analyze

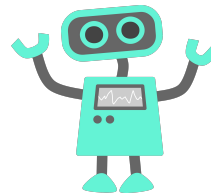
Results



Annif rajapintapalveluna

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

Suggest subjects!



Annif API

```
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"}
]
```

Algoritmeja automaattiseen sisällönkuvailuun

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](https://doi.org/10.1007/s00799-018-0240-3)

Algoritmeja voi käyttää **yksitellen** tai **yhdistelminä** (ensemble)



Maui, TFIDF,
FastText, vw_multi ...



ensemble, pav, vw_ensemble,
nn_ensemble ...

Mitä Annifilla voi tehdä?

Jyväskylän yliopiston JYX-julkaisuarkisto

Opiskelijat lataavat omat gradunsa, Annif ehdottaa niille aiheita

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

Wikipedian kuvailu YSOlla

Suomenkielisessä Wikipediassa on 410 000 artikkelia (raakatekstinä 620 MB)

Automaattinen sisällönkuvailu kesti n. 6-7 tuntia

Aiheita annettiin 1-3 per artikkeli (keskimäärin n. 2)

Esimerkkejä: (satunnaisesti valittu)

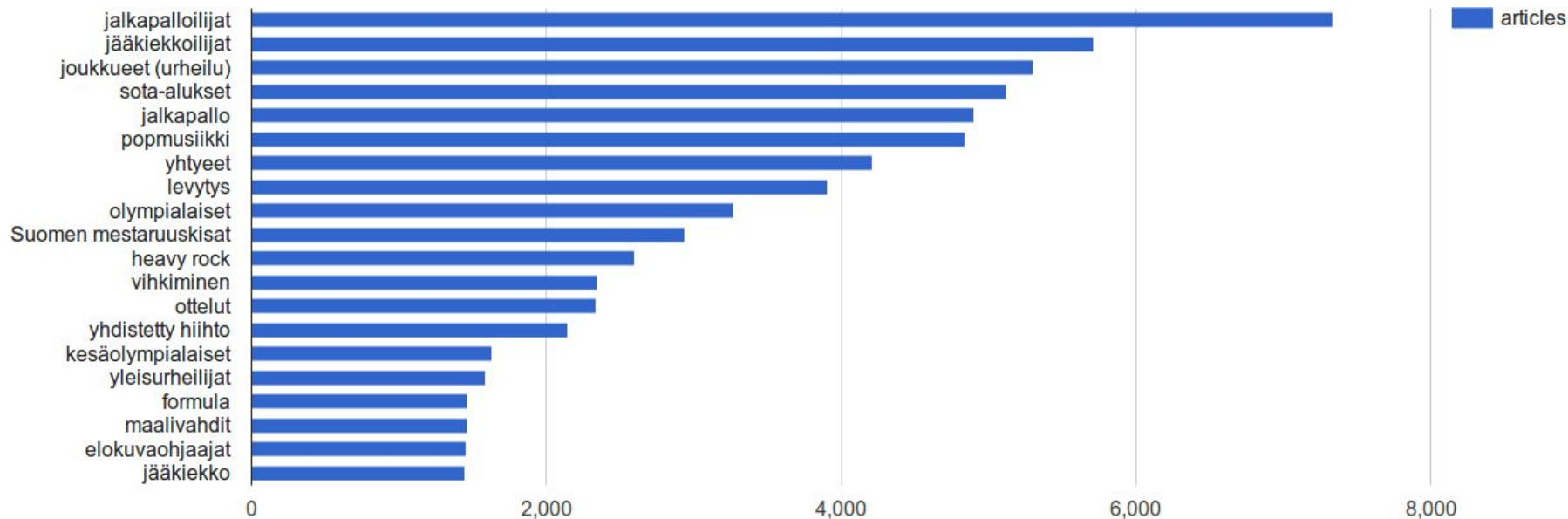
Wikipedia-artikkeli

Ahvenuslammi (Urjala)
Brasilian Grand Prix 2016
Guy Topelius
HMS Laforey
Liigacup
Pää Kii
RT-21M Pioneer
Runoja
Sjur Røthe
Veikko Lavi

YSO-aiheet

rannat
kilpa-autoilijat, formula, **mikroautoilu**
kansanrunoudentutkijat, **sakariini**
sota-alukset
jalkapallo, jalkapalloilijat
yhtyeet, popmusiikki
ohjukset
popmusiikki, levytys, sävellykset
hiihtäjät, hiihto, yhdistetty hiihto
sanoittajat, kupletit

Yleisimmät aiheet Wikipediassa



Yleisimmät aiheet Wikipediassa



Image credits:

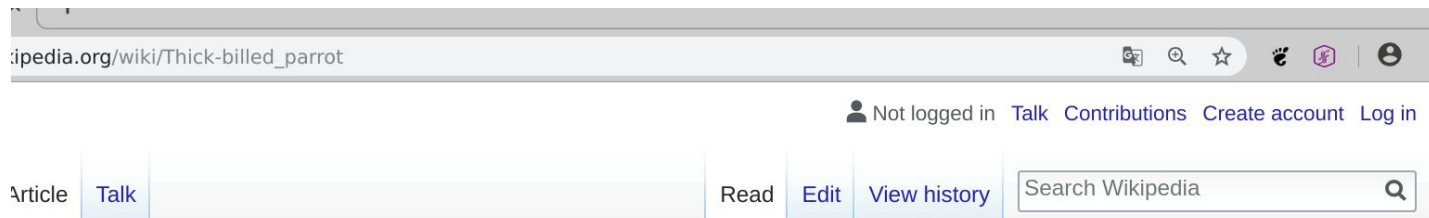
Petteri Lehtonen [CC BY-SA 3.0]

Hockeybroad/Cheryl Adams [CC BY-SA 3.0]

Tomisti [CC BY-SA 3.0]

Tuomas Vitikainen [CC BY-SA 3.0]

Finna Recommends -selainlaajennos Chromelle



Thick-billed parrot

From Wikipedia, the free encyclopedia

This page is about the species of parrot. For the genus of parrots, see [Rhynchopsitta](#).

The **thick-billed parrot** (*Rhynchopsitta pachyrhyncha*) is a medium-sized green and red [parrot](#) found in Mexico, that formerly ranged into the [southwestern United States](#). Its position in parrot [phylogeny](#) is the subject of ongoing discussion; it is sometimes referred to as thick-billed [macaw](#) or thick-billed [conure](#). In Mexico, it is locally called *guacamaya* ("macaw") or *cotorra serrana* ("mountain parrot"). Classified internationally as Endangered through [IUCN](#),^[1] the thick-billed parrot's decline has been central to multiple controversies over wildlife management.

Contents [\[hide\]](#)

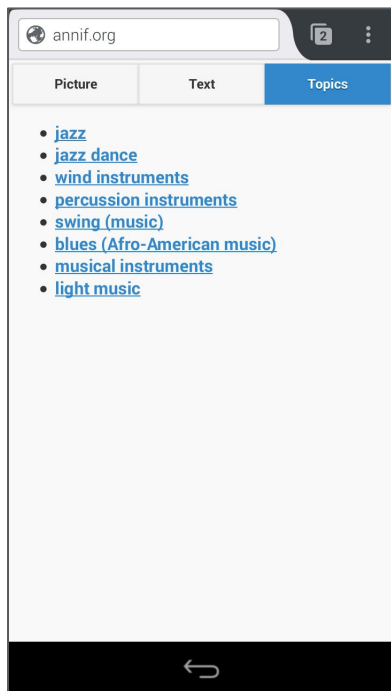
- [Taxonomy](#)
- [Description](#)



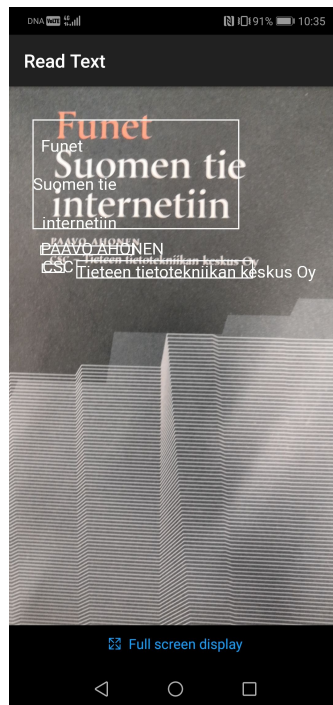
Analysoi
valitun tekstin
miltä tahansa sivulta
Annifin APilla
ja hakee
kirjasuosituksia
Finnasta

WIDE-hackathon
Yazan Alhalabi
Samuel Akangbe
Steven Nebo

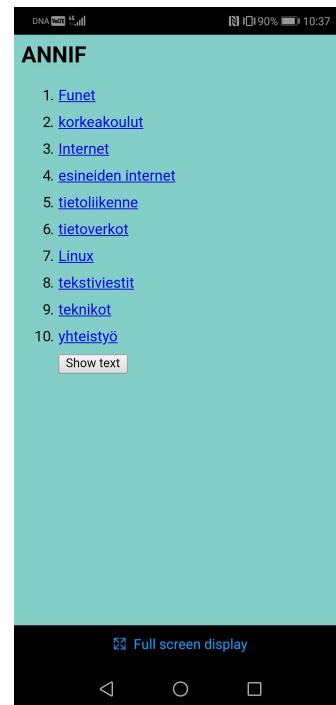
Mobiilisovellukset



Prototyyppi web-sovellus
OCR-pilvipalvelu ocr.space
m.annif.org



Prototyyppi Android-sovelluksesta, OCR laitteessa
(tekijä Okko Vainonen)



bot.annif.org

Annifia tekoälynä käytävä chatbot,
joka suosittelee kuvia ja kirjoja Finnasta

Tervetuloa juttelemaan!



Anni F.

Hei, olen Anni! Hauska tutustua!

Mikä sinua tänään kiinnostaa?

Kiinnostaako sinua velhot?

Kiinnostaako sinua fantasia?

Odotas, etsin sinulle kuvan!

HS Ask 9.16.1



Haluatko lisää?

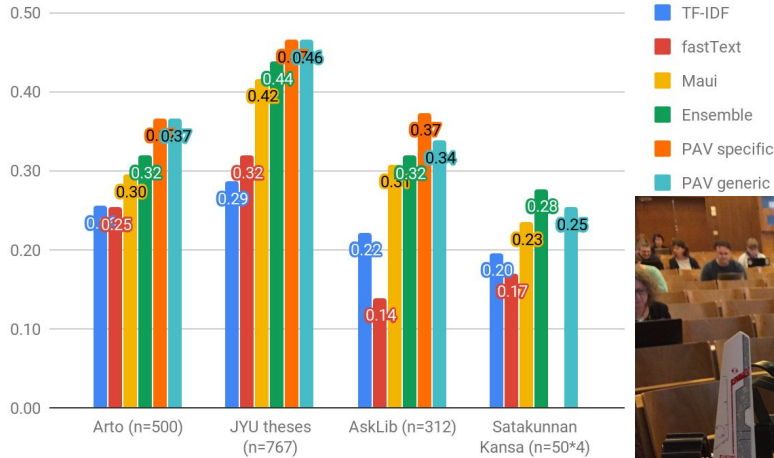
Harry Potter

Ei

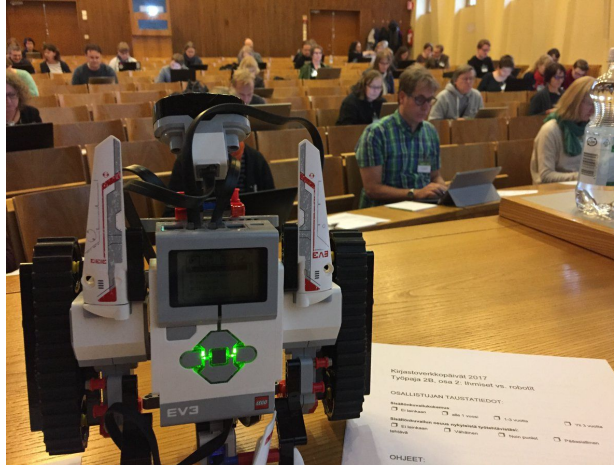
Joo!

**Neljä tavoitetta
lähitulevaisuudelle**

1. Laadun parantaminen



Algoritmien testaus ja arviointi yhdessä CSC:n kanssa

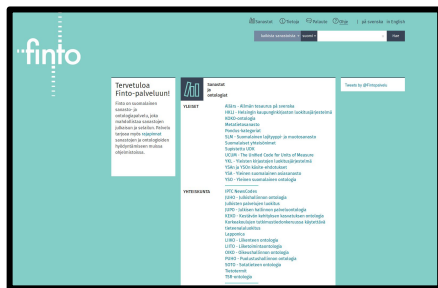


Ihmiset vs. robotit työpaja
Kirjastoverkkopäivillä 2017

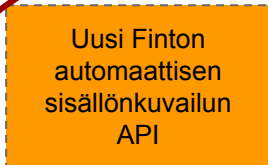


Aaveita koneessa -työpaja
Kirjastoverkkopäivillä 2019

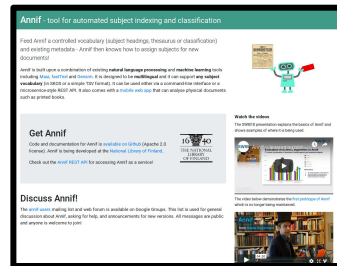
2. Tuotantokäyttöön sopiva rajapintapalvelu



api.finto.fi/rest/
rajapinta sanastodataan
tuotantokäyttöä varten

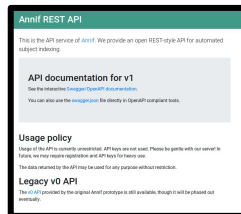


api.annif.finto.fi
rajapinta autom. kuvailuun
tuotantokäyttöä varten



annif.org

Annif-ohjelmiston käyntikorttisivusto



api.annif.org
demorajapinta automaattiseen
sisällönkuvailuun

Finton rajapinnan laajentaminen automaattisen sisällönkuvailun palveluihin Annifin avulla

3. Käyttöönotto Kansalliskirjaston järjestelmissä

The screenshot displays the LUT University website with three main digital repository interfaces:

- Doria:** The top left interface, featuring a search bar and a description: "Doria on Kansalliskirjaston ylläpitämä julkaisuarkisto, josta löydät laajan valikoiman julkaisuja." It includes a link to "Lisää tietoa Doriasta".
- LUTPub:** The top right interface, described as "LUTPub on LUT-yliopiston avoin julkaisuarkisto." It mentions that it contains LUT-university's theses, diplomas, and research articles, and provides a link to "Lisää tietoa LUTPubista".
- Julkari:** The bottom left interface, described as "Julkari on Sosiaali- ja terveysministeriön julkaisuarkisto." It states that it contains the latest publications from the Ministry of Social and Health Services and provides a link to "Lisää tietoa Julkari".

The central part of the screenshot shows the "Theseus" interface, which is the LUT University's open access repository for theses and dissertations. It includes a search bar and a description: "Theseus - ammattikorkeakoulujen opinnäytetyöt ja julkaisut verkossa".

The screenshot shows the "Digitaalisten julkaisujen arkistointi" (Digital publication archiving) form from the Kansalliskirjasto. The form is divided into several sections:

- Luovuttajan tiedot:** Fields for "Yhteystiedot" (Contact information), "Sähköpostiosoite" (Email address), "Puhelinnumero" (Phone number), and "Organisaatio" (Organization). There is a checkbox for "Muuta näitä tietoja" (Other information).
- Julkaisun tiedot:** Fields for "Julkaisun nimi" (Publication name), "Julkaisun tyyppi" (Publication type), and "Päättämistiedot" (Finalization information).
- ISBN (viiroja):** A field for the ISBN number.

At the bottom, there is a section for "E-vapaakappaleiden vastaanotto ja käsittely" (Acceptance and processing of e-copies).

DSpace-pohjaiset julkaisuarkistot

4. Kansainvälinen käyttäjäyhteisö

Annif is a multi-algorithm automated subject indexing tool for libraries, archives and museums. This repository is used for developing a production version of the system, based on ideas from the initial prototype. <http://annif.org>

[subject-indexing](#) [python](#) [machine-learning](#) [code4lib](#) [classification](#) [rest-api](#) [flask-application](#) [contexton](#)

1,104 commits 15 branches 52 releases 6 contributors View license

Branch: master New pull request

Find file Clone or download

 osma Merge pull request #336 from NatLuf/156f-optimizations	Latest commit 462165f 16 days ago
 annif Cleanup unused imports	16 days ago
 docs fix RTD-build-error	4 months ago
 tests Perform document to subject conversion in memory; remove stale Subject...	19 days ago
 codeclimate.yml more comprehensive Code Climate configuration	2 years ago
 codecov.yml Codecov should ignore setup.py	2 years ago
 coveragerc Generate Codecov reports	2 years ago
 dockerignore Split context copying to avoid rebuilding pipenv layer on code change	
 drone.yml Switch trigger to listen to master	
 gltignore Git-ignore Sphinx build html documentation	
 lgfm.yml Add LGTM configuration excluding fasttext	
 readthedocs.yml Install sphinxcontrib-apidoc via documentation specific requirements.txt	

Annif Users Jaettu julkisesti
30 useista aiheista

Welcome to the [Annif users'](#) mailing list / web forum! This list can be used for

- general discussion about Annif, its features and usage scenarios
- asking for help with installing or running Annif
- future directions for Annif
- announcements for new versions and other Annif-related news

The list is open for anyone to join and all messages are public. You need to join the list yourself before posting.

 Segmentation fault (core dumped) error when trying to train PAV Tekijä: Pekka Kauranen - 1 viesti - 2 katselukertaa	18. lokakuuta
 Annif questionnaire Tekijä: Mona Lehtinen - 6 viestiä - 48 katselukertaa	26. syyskuuta
 skos format Tekijä: Sara Veldhoen - 3 viestiä - 14 katselukertaa	20. syyskuuta
 which elements of skos vocabulary are used in matching? Tekijä: Enrico - 8 viestiä - 23 katselukertaa	19. syyskuuta
 ANN: Annif 0.42, with enhancements to CLI, more robust handling of inputs, versioning of Docker image... Tekijä: Juho Inkinen - 1 viesti - 8 katselukertaa	2. syyskuuta

Koodi, kehitys ja dokumentaatio GitHubissa

Annif-users käyttäjäfoorumi



Johdatus Annifin käyttöön SWIB19-konferenssissa Hampuri 25.11.2019

Automaattisen kuvailun verkosto

Lisännyt Mikko Lappalainen, viimeksi muokannut Mona Lehtinen kesäkuuta 24, 2019

Automaattisen kuvailun verkosto on avoin verkosto, jonka tavoitteena on jakaa tietoa erilaisista kuvailun automatisointiin liittyvistä ratkaisuksista. Verkosto kokoontuu tarpeen mukaan muutaman kerran vuodessa.

Verkostolle on myös perustettu s-postilista, jota tullaan käyttämään lähinnä tiedottamiseen. Listalle saa liittyä vapaasti lähettämällä sähköpostia osoitteeseen majordomo (ät) helsinki.fi. Huomaa, että:

1. Viestin otsikkokentän tulee olla tyhjä
2. Tekstiosassa tulee lukea pelkästään: subscribe automaattinen-kuvailu (Mahdollinen allekirjoituskin kannattaa siis poistaa viestistä.)

Listalta voi erota vastaavasti lähettämällä em. osoitteeseen tyhjällä otsikolla viesti unsubscribe automaattinen-kuvailu.

<https://www.kiwi.fi/display/tekoalykumppanuus/Automaattisen+kuvailun+verkosto>

Kiitos!

osma.suominen@helsinki.fi - [@OsmaSuominen](https://twitter.com/OsmaSuominen)

<http://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>

Tämä esitys: <https://tinyurl.com/annif-bmf>