

Automaattisen kuvailun työkalu Annif: tulevaisuuden näkymiä

Osma Suominen

Kirjastoverkkopäivät 2019
23.10.2019

Extrablad till ÅBO UNDERRÄTTELSE

No 2.
Finlands oavhängighet.

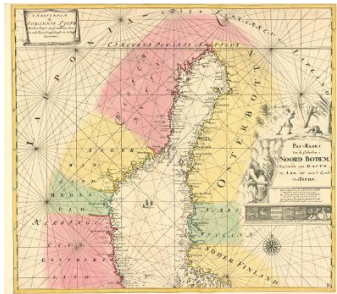
Landsdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudsaken i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 rösterna mot 58 vilka till följde ett av socialdemokraternas förordnad förslag.

Verkligheten beträffande
Finland har varit en del af det svenska riket sedan 1809. Sedan 1809 till 1812 har Finland varit en del af det svenska riket. Sedan 1812 har Finland varit en del af det svenska riket. Sedan 1812 har Finland varit en del af det svenska riket.

Årskiftet 1809-1812
Årskiftet 1809-1812 har varit ett av de mest betydande i Finlands historia. Det har varit ett av de mest betydande i Finlands historia. Det har varit ett av de mest betydande i Finlands historia.

Årskiftet 1812-1818
Årskiftet 1812-1818 har varit ett av de mest betydande i Finlands historia. Det har varit ett av de mest betydande i Finlands historia. Det har varit ett av de mest betydande i Finlands historia.

Pris 25 penn.



JANIS 1858
IN 1870

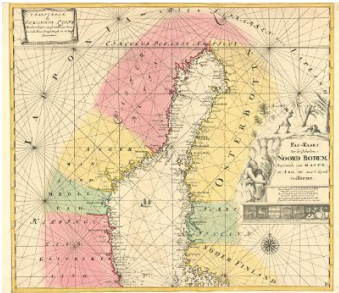
Framework for Open Science and Research 3.2.2019 6.1916

For example, which physical disciplines will store the digital data resources, who will be responsible using them for advanced research systems?

The Kartus EA method used in the higher education sector (described below) has been explained in this work.

Open Level	Open Level	Open Level	Open Level	Open Level
Level 1: Open Access	Level 2: Open Access	Level 3: Open Access	Level 4: Open Access	Level 5: Open Access
Level 6: Open Access	Level 7: Open Access	Level 8: Open Access	Level 9: Open Access	Level 10: Open Access
Level 11: Open Access	Level 12: Open Access	Level 13: Open Access	Level 14: Open Access	Level 15: Open Access
Level 16: Open Access	Level 17: Open Access	Level 18: Open Access	Level 19: Open Access	Level 20: Open Access
Level 21: Open Access	Level 22: Open Access	Level 23: Open Access	Level 24: Open Access	Level 25: Open Access
Level 26: Open Access	Level 27: Open Access	Level 28: Open Access	Level 29: Open Access	Level 30: Open Access
Level 31: Open Access	Level 32: Open Access	Level 33: Open Access	Level 34: Open Access	Level 35: Open Access
Level 36: Open Access	Level 37: Open Access	Level 38: Open Access	Level 39: Open Access	Level 40: Open Access
Level 41: Open Access	Level 42: Open Access	Level 43: Open Access	Level 44: Open Access	Level 45: Open Access
Level 46: Open Access	Level 47: Open Access	Level 48: Open Access	Level 49: Open Access	Level 50: Open Access

In accordance with the Kartus EA method, advanced work was conducted by the Kartus EA EA description model by using the physical disciplines for open science. As the objective of the enterprise architecture description is the right work for open science, the focus has been on the activities that describe the objectives of the digital open science, in accordance with the Kartus model. The following diagram roughly models the sub-description of the enterprise architecture conducted in efforts.



Framework for Open Science and Research

1.1.2019

8.196

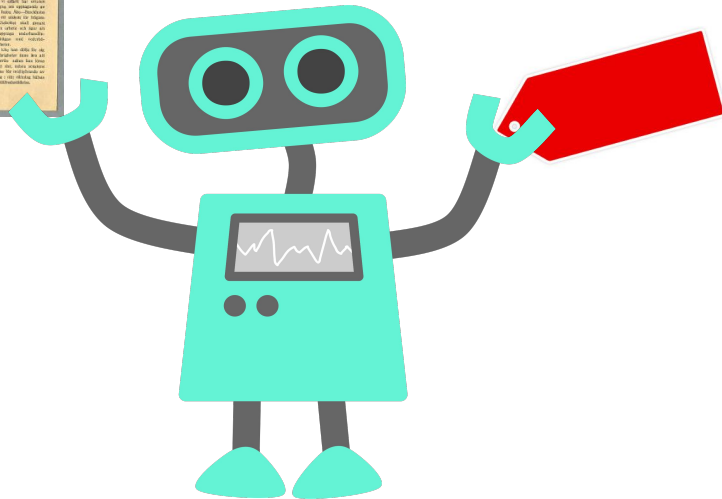
For example, which physical disciplines will use the digital data resources, who will be implementing using data for advanced measurement systems?

The Karsten EA method used in the higher education sector (described below) has been explained in this work.

Open Level	Level	Level	Level	Level
Open Level	Level	Level	Level	Level
Open Level	Level	Level	Level	Level
Open Level	Level	Level	Level	Level
Open Level	Level	Level	Level	Level

In accordance with the Karsten EA method, advanced work was conducted by the Karsten EA description model based on the Karsten EA description model for open science. As the objective of the enterprise architecture description is the right structure work for open science, the focus has been on the activities that describe the objectives of the digital open science, in accordance with the Karsten model. The following diagram roughly models the sub-description of the enterprise architecture conducted in efforts.





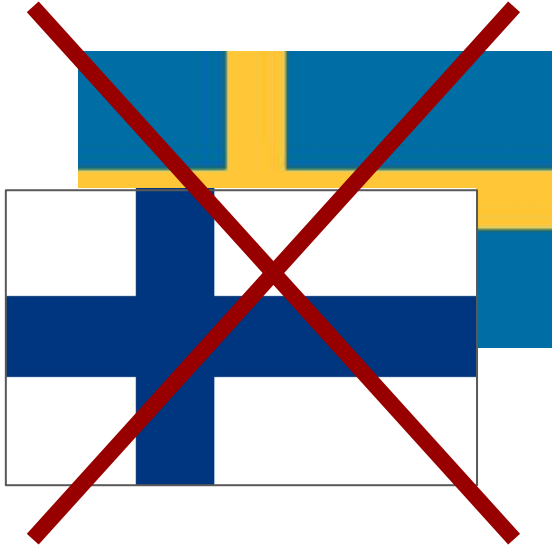


OPEN
CALAIS



THOMSON REUTERS





~~YSA YSO
Allärs KOKO~~

€ £ \$

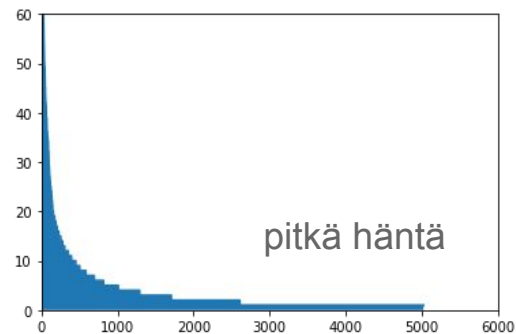
Sisällönkuvailu on vaikeaa

ihmisille:

- **Subjektiiivisuus:** 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



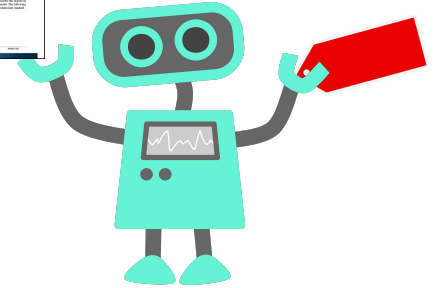
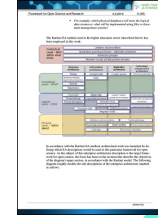
Annif

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 haulla.

Hae...

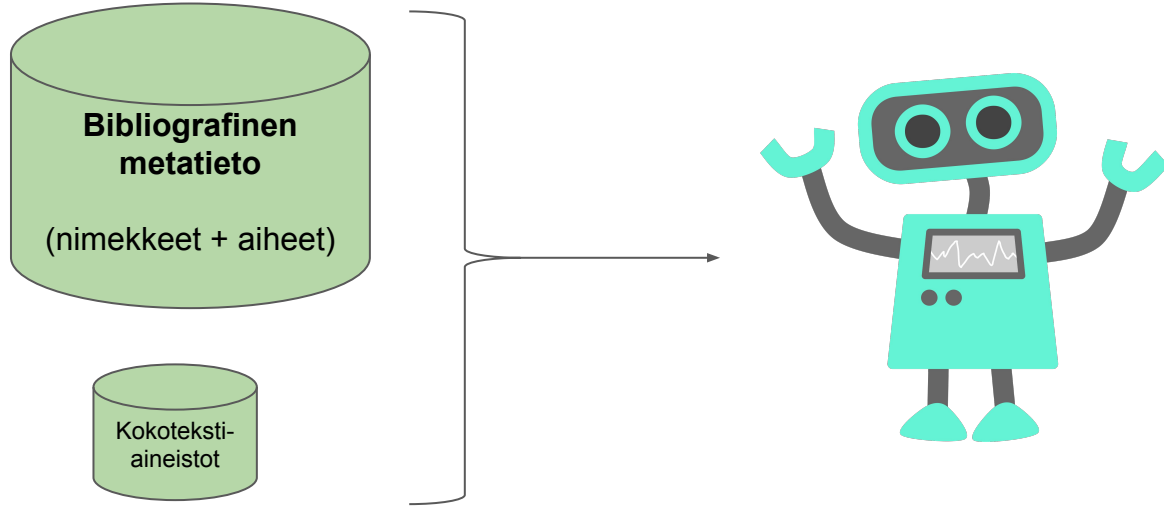
Kaikki osumat ▾



🔍 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 suojaautua

Project (vocabulary and language):

YSO ensemble Finnish

Analyze

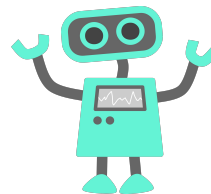
Results



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.”

ys0:p14146
“renewable natural resources”

Associative approaches

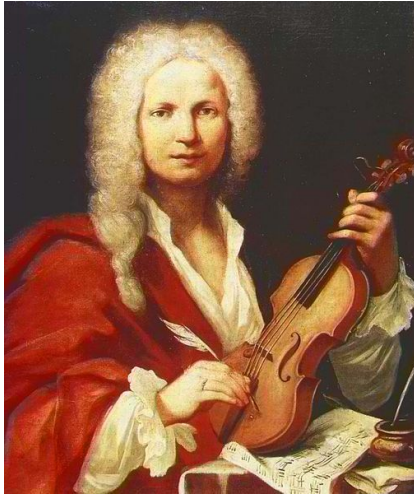
Learn which **concepts** are correlated with which **terms** in documents, based on training data



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)

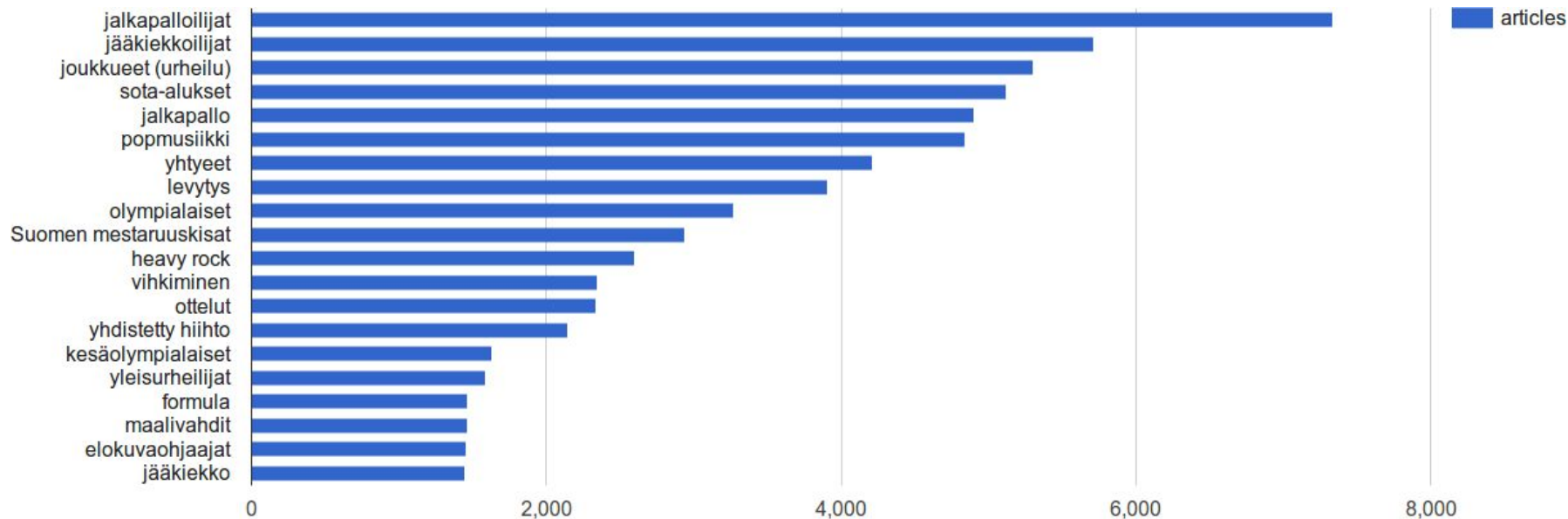
Wikipedia-artikkeli

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:

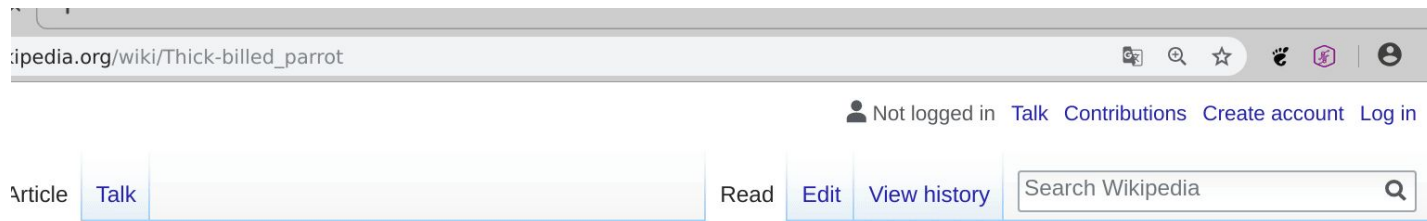
Petteri Lehtonen [CC BY-SA 3.0]

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 [\[hide\]](#)

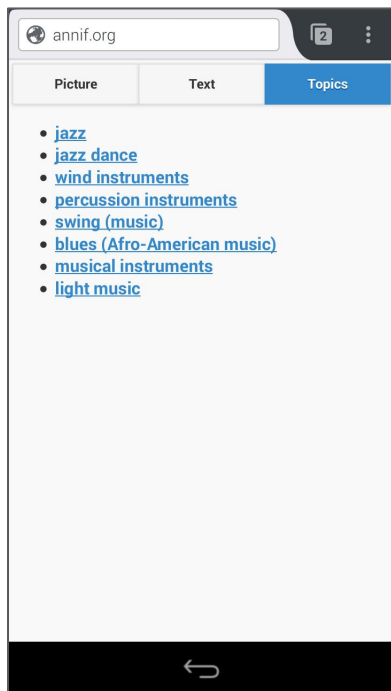
- [Taxonomy](#)
- [Description](#)



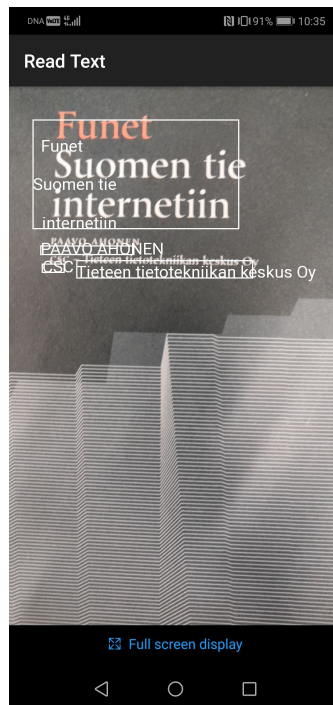
Analysoi
valitun tekstin
miltä tahansa sivulta
Annifin APilla
ja hakee
kirjasuosituksia
Finnasta

WIDE-hackathon
Yazan Alhalabi
Samuel Akangbe
Steven Nebo

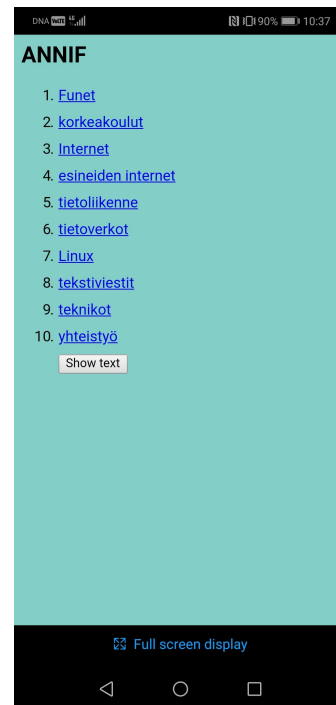
Mobiilisovellukset



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



KIRJASTOVERKKOPÄIVÄT

Helsinki 23.-24.10.2019

pala palalta parempi



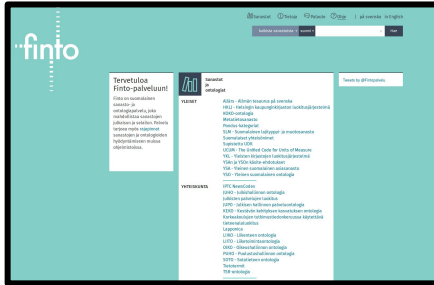
Algoritmien testaus ja arviointi yhdessä CSC:n kanssa



Ihmiset vs. robotit työpaja
Kirjastoverkkopäivillä 2017

Työpajan nimi	Työpaja 4 Automaattinen sisällönkuvaluu: Aaveita koneessa
Kellonaika	9.30–12.00
Paikka	
Yhteyshenkilö	Osma Suominen, Kansalliskirjasto
Työpajan kuvaus	Työpajassa tarkastellaan sisällönkuvilun laatua eri näkökulmista. Käytännön osuudessa arvioidaan ja pisteytetään esimerkiksi kirjjoille tehtyjä sisällönkuviluja. Tavoitteena on muodostaa yhteinen käsitys siitä, mitä on laadukas sisällönkuvaluu sekä vertailla, miten koneellisesti tai koneavusteisesti tuotetut sisällönkuvilut eroavat ihmistyönä tehdyistä. Työpaja on jatkoa vuoden 2017 Kirjastoverkkopäivien Ihmiset vs. robotit -työpajalle.
Kohderyhmä	Sisällönkuvilusta kiinnostuneet
Tietoa osallistujille työpajan työskentelyyn ja tarvikkeisiin liittyen	Alustukset n. 30 min, jonka jälkeen vuorossa on käytännön harjoitusten tekemistä omalla läppärillä selainpohjaisessa ympäristössä. Lopuksi tulokset analysoidaan ja tehdään niistä yhteenveto. Tabletilla osallistuminen voi olla vaikeaa, joten emme suosittele sitä. Muutama varakone on tarjolla, jos et jostain syystä voi tuoda omaa tietokonetta. Varmista myös, että akku on täynnä.

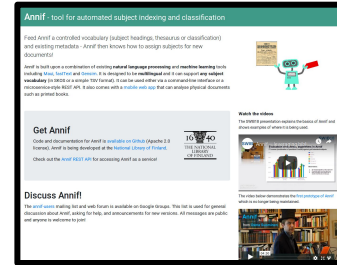
2. Tuotantokäyttöön sopiva rajapintapalvelu



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

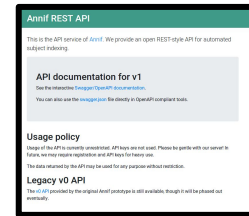
Uusi Finton
automaattisen
sisällönkuvailun
API

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 image shows two overlapping screenshots of digital library search interfaces. The top screenshot is the LUT University Doria interface, featuring a search bar with the text 'LUTPubin haku- ja käyttöohje' and a 'HAE' button. Below it, there's a section for 'LUTPub' with a description: 'LUTPub on LUT-yliopiston avoin julkaisuarkisto. LUTPub sisältää LUT-yliopiston kandidaatintöitä ja -tutkielmia, diplomitöitä, pro gradu -tutkielmia, lisensiaattintöitä, väitöskirjoja sekä näiden tiivistelmiä. Lisäksi LUTPubiin on tallennettu artikkeleita, tutkimusaineistoja ja esitysmateriaaleja.' The bottom screenshot is the Julkari interface, with a search bar containing 'Lauda' and a 'Haku' button. It includes a description: 'Julkari on Sosiaali- ja terveysministeriön julkaisuista on saatavilla sähköinen versi laitoksen osalta myös julkaisurekisterinä. Julkariin julkaisuja voi hakea mm. julkaisjärjestetty kokoelmittain, joten sanahaun avulla voit löytää kaikki julkaisut, jotka sisältävät lausekkeen 'Lauda'. Julkari on vapaasti käytettävissä. Julkaisuarkisto on saatavilla myös mobiililaitteilla. Lisää tietoa Julkariin.' Both screenshots show navigation menus and user profile icons.

The image shows a screenshot of the National Library of Finland's digital publication archiving interface. The header reads 'Digitaalisten julkaisujen arkistointi'. Below the header, there are several sections for metadata entry: 'Luovuttajan tiedot', 'Julkaisun tiedot', and 'Julkaisun tekemä'. The 'Luovuttajan tiedot' section includes fields for 'Yhteystietokäyttö', 'Sähköpostiosoite', 'Puhelinnumero', and 'Organisaatio'. The 'Julkaisun tiedot' section includes a 'Julkaisun tekemä' field with a dropdown menu and a 'Julkaisun tyyppi' field with radio buttons for 'käsikirja', 'muut', 'äänite', and 'muu'. The 'Julkaisun tiedot' section also includes a 'Päättämätiedot' field with an 'ISBN (vuvuilla)' label. The interface is clean and professional, with a dark header and light background.

E-vapaakappaleiden vastaanotto ja käsittely

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 comexon

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

Branch: master New pull request Find file Clone or download

Author	Commit Message	Latest commit
oasma	Merge pull request #336 from NatLibFi/tdf-optimizations	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	
gignore	Git-ignore Sphinx build html documentation	
lgfm.yml	Add LGTM configuration excluding fasttext	
readthedocs.yml	Install sphinxcontrib-apidoc via documentation specific requirements.txt	

Koodi, kehitys ja dokumentaatio GitHubissa



Johdatus Annifin käyttöön SWIB19-konferenssissa Hampuri 25.11.2019

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

Annif-users käyttäjäfoorumi

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 ratkaisuista. 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-kivepa2019>