Automaattisen kuvailun työkalu Annif: tulevaisuuden näkymiä

Osma Suominen

Kirjastoverkkopäivät 2019
23.10.2019

Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

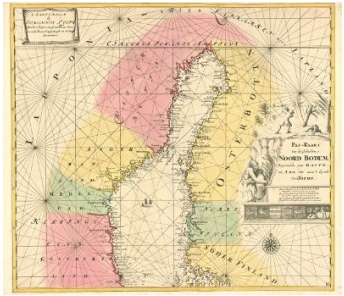
Finlands oavhängighet.

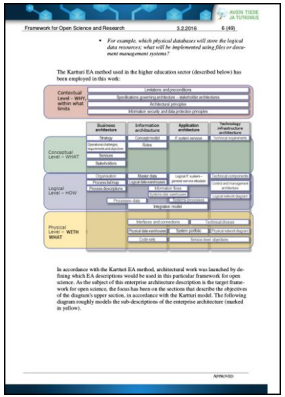
Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.

Våldsdåd av arbetare.

Återupptagandet av trafiken Åbo—Stockholm.

VENNY SOLDAN-BROFELDT.





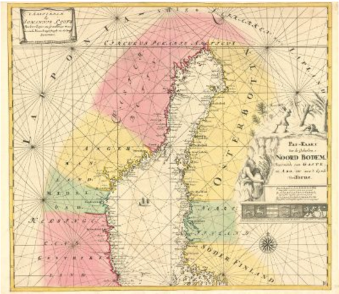
Framework for Open Science and Research 3.2.2016 6 (49)

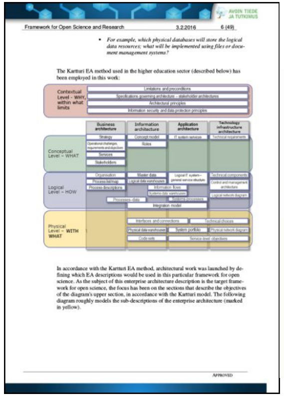
- *For example, which physical databases will store the logical data resources; what will be implemented using files or document management systems?*

The Kartturi EA method used in the higher education sector (described below) has been employed in this work:

In accordance with the Kartturi EA method, architectural work was launched by defining which EA descriptions would be used in this particular framework for open science. As the subject of this enterprise architecture description is the target framework for open science, the focus has been on the sections that describe the objectives of the diagram's upper section, in accordance with the Kartturi model. The following diagram roughly models the sub-descriptions of the enterprise architecture (marked in yellow).

Extrablad till
ÅBO UNDERRÄTTELSER
N:o 2.
Finlands oavhängighet.
Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.
Våldsdåd av arbetare.
Återupptagandet av trafiken Åbo—Stockholm.
Pris 25 penni.





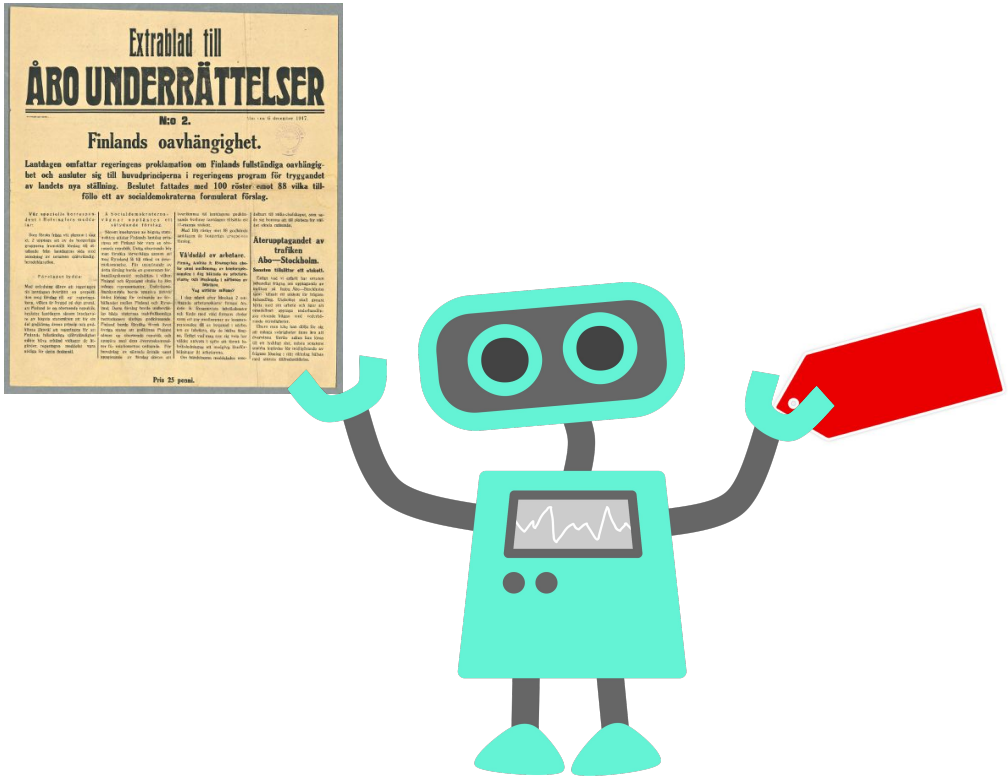
AVOIN TIEDE JA TUTKIMUS
Framework for Open Science and Research 3.2.2016 6 (49)

- *For example, which physical databases will store the logical data resources; what will be implemented using files or document management systems?*

The Kartturi EA method used in the higher education sector (described below) has been employed in this work:

In accordance with the Kartturi EA method, architectural work was launched by defining which EA descriptions would be used in this particular framework for open science. As the subject of this enterprise architecture description is the target framework for open science, the focus has been on the sections that describe the objectives of the diagram's upper section, in accordance with the Kartturi model. The following diagram roughly models the sub-descriptions of the enterprise architecture (marked in yellow).

APPROVED



Extrablad till

ÅBO UNDERRÄTTELSER

N:o 2.

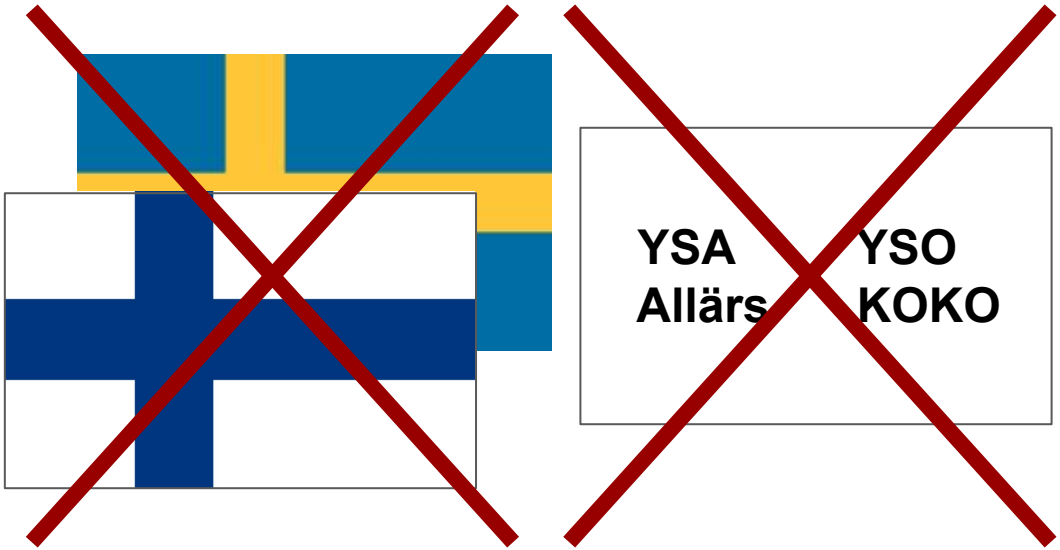
Finlands oavhängighet.

Lantdagen omfattar regeringens proklamation om Finlands fullständiga oavhängighet och ansluter sig till huvudprinciperna i regeringens program för tryggandet av landets nya ställning. Beslutet fattades med 100 röster emot 88 vilka tillföllo ett av socialdemokraterna formulerat förslag.

Pris 25 penni.

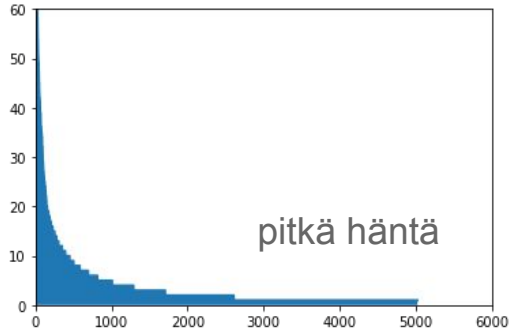
OPEN CALAIS

THOMSON REUTERS


YSA
Allärs
YSO
KOKO

€£$

Sisällönkuvailu on vaikeaa



ihmisille:

- **Subjektiivisuus**: kun kaksi eri ihmistä kuvailee saman dokumentin, vain ~⅓ 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

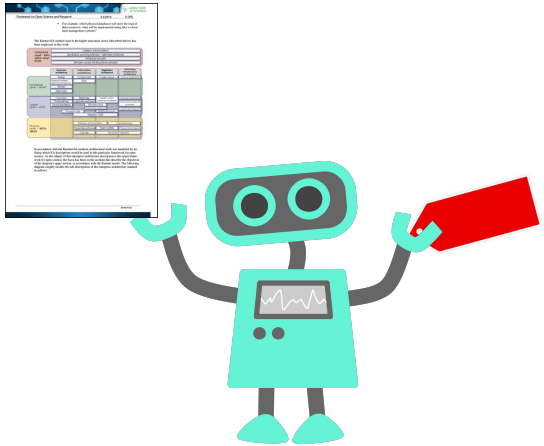
Annif

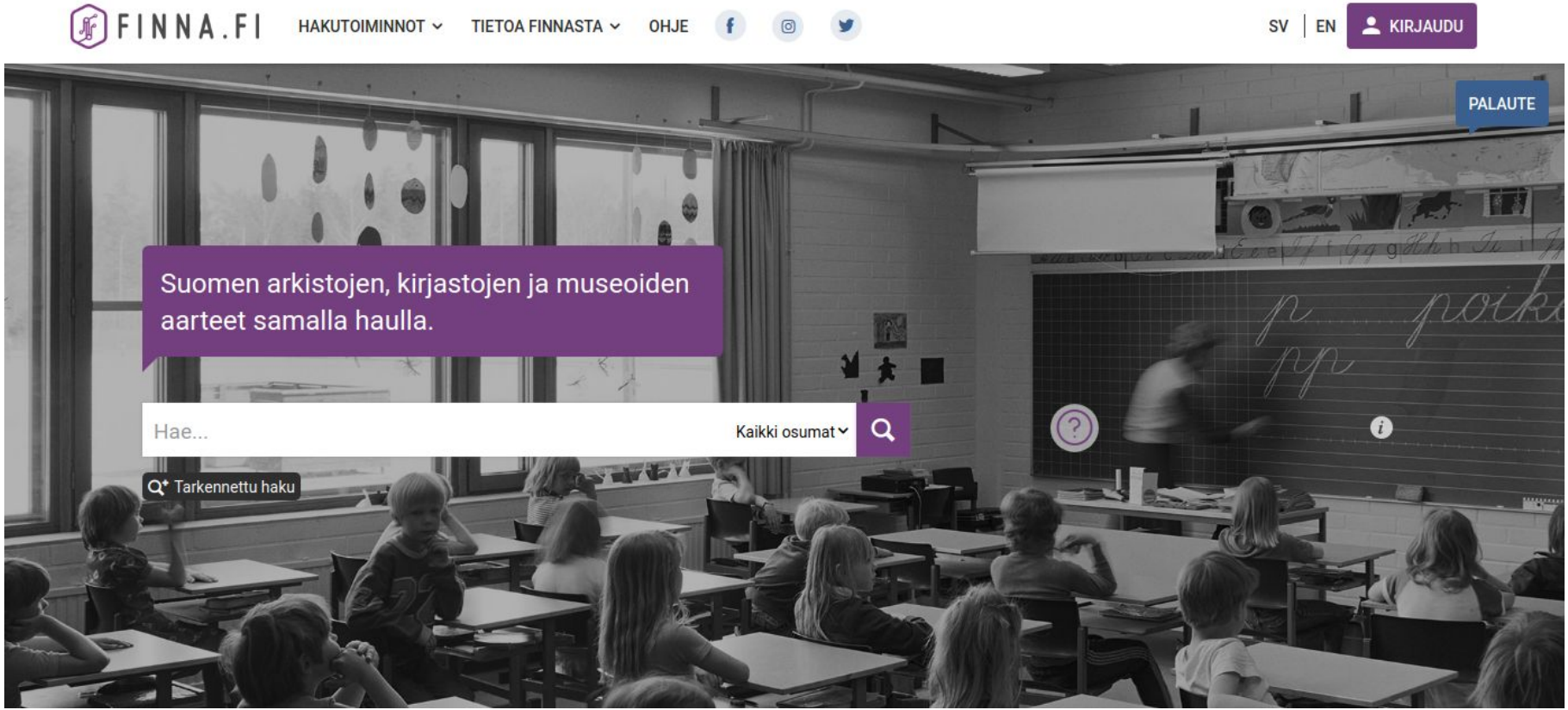
Tuottaa sisällönkuvailua (asiasanoitus ja luokitus) tekstin perusteella

Kansalliskirjastossa kehitetty työkalu, avointa koodia

Soveltuu parhaiten asiatekstille, mutta testataan myös muilla aineistoilla

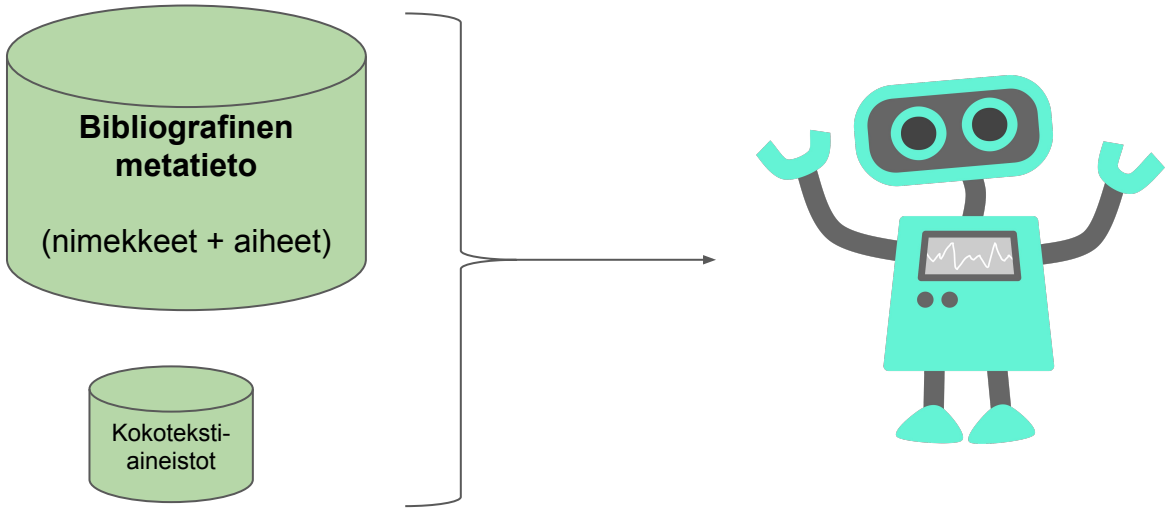
Perustuu **kieliteknologiaan** ja **koneoppimiseen**





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

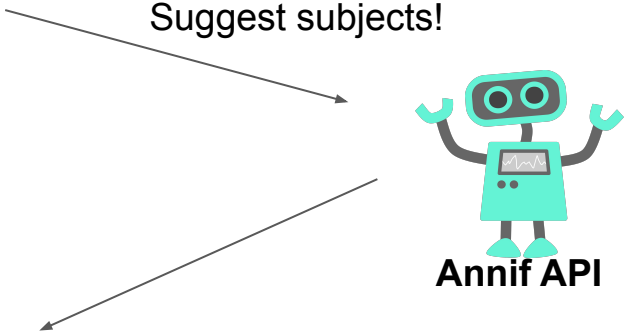
Analyze

Results

- yksinäisyys
- viestimet
- kasvot
- väitöskirjat
- maisterit
- eleet
- reaktiot
- katse
- ilmeet
- mieliala

Annif rajapintapalveluna

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



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

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

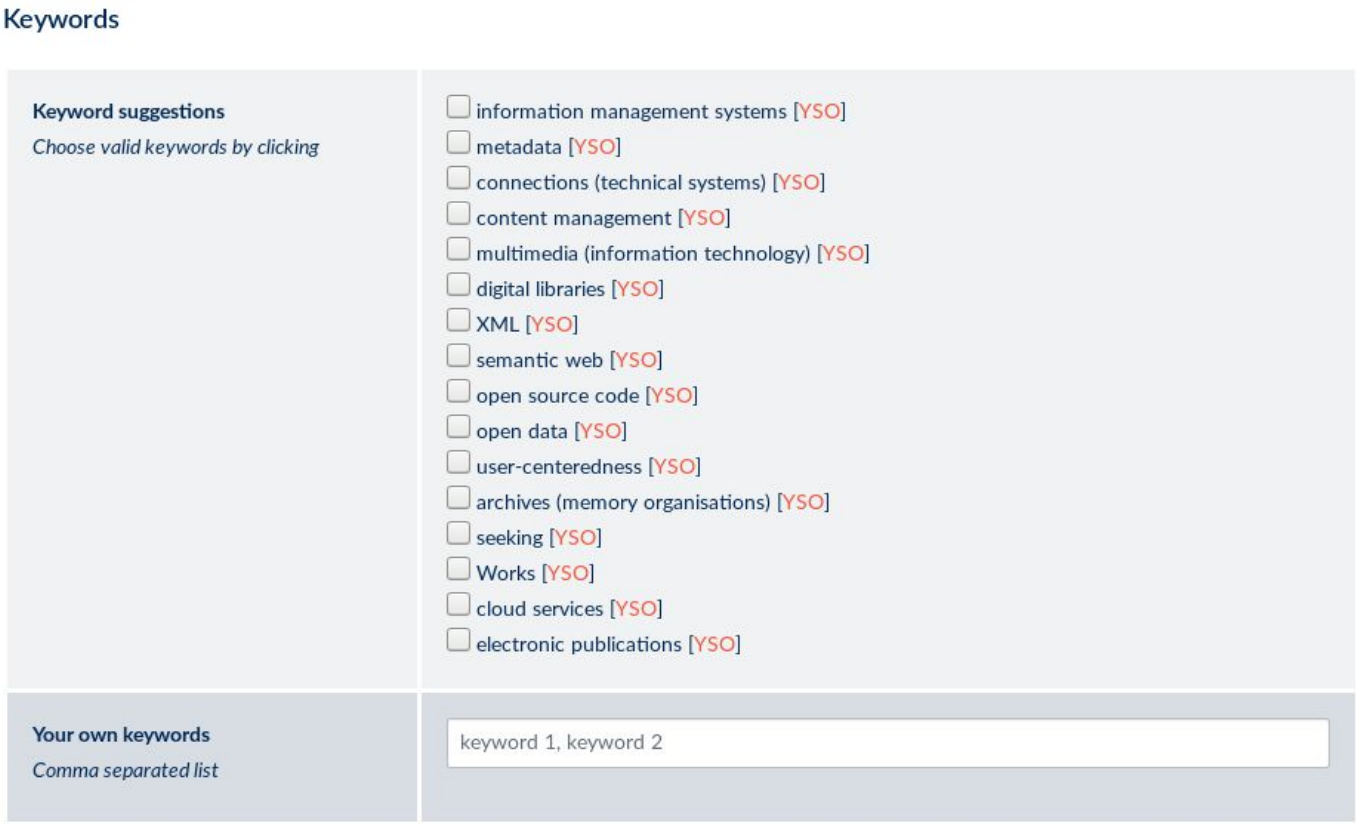
Maui, TFIDF,
FastText, vw_multi ...

ensemble, pav, vw_ensemble,
nn_ensemble ...

Mitä Annifilla voi tehdä?

Jyväskylän yliopiston JYX-julkaisuarkisto

Opiskelijat lataavat omat gradunsa, Annif ehdottaa niille aiheita



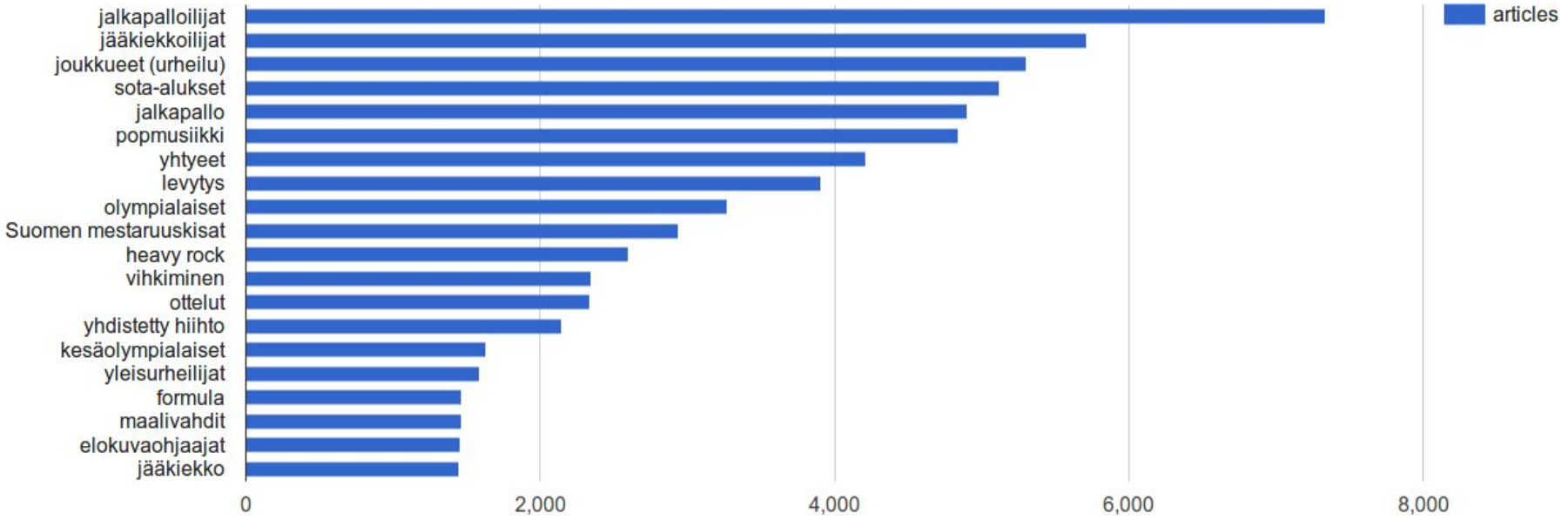
Wikipedian kuvailu YSOlla

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	**YSO-aiheet**
Ahvenuslammi (Urjala)	rannat
Brasilian Grand Prix 2016	kilpa-autoilijat, formula, mikroautoilu
Guy Topelius	kansanrunoudentutkijat, sakariini
HMS Laforey	sota-alukset
Liigacup	jalkapallo, jalkapalloilijat
Pää Kii	yhtyeet, popmusiikki
RT-21M Pioneer	ohjukset
Runoja	popmusiikki, levytys, sävellykset
Sjur Røthe	hiihtäjät, hiihto, yhdistetty hiihto
Veikko Lavi	sanoittajat, kupletit

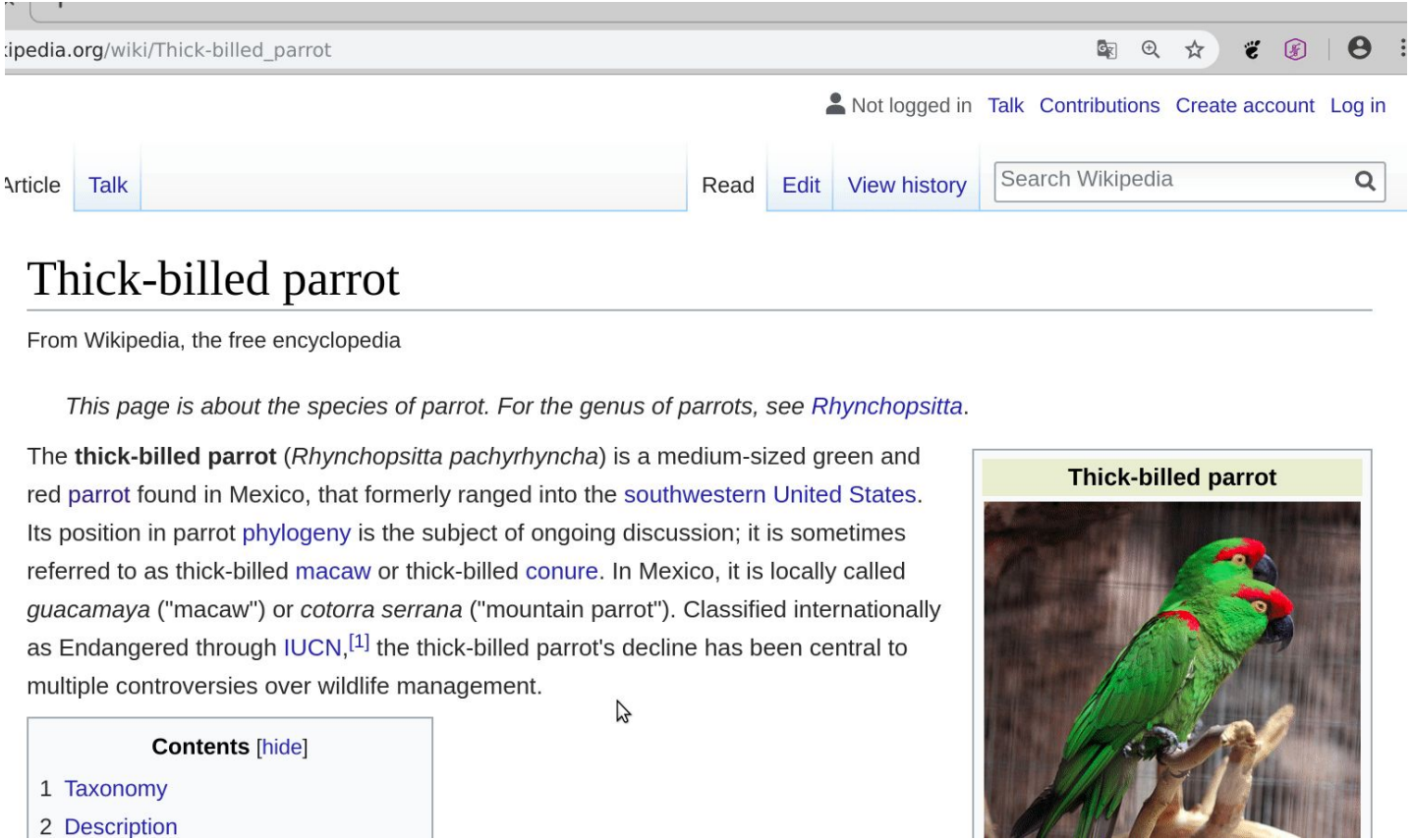
Yleisimmät aiheet Wikipediassa



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]

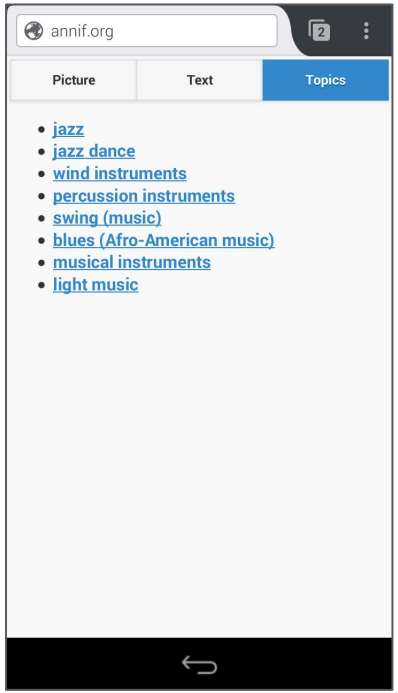
Finna Recommends -selainlaajennos Chromelle



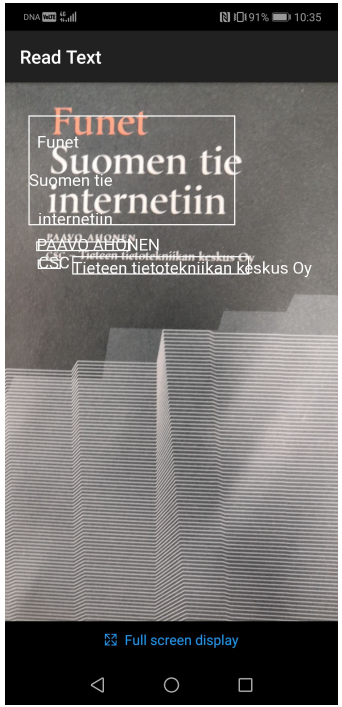
Analysoi
valitun tekstin
miltä tahansa sivulta
Annifin APIlla
ja hakee
kirjasuosituksia
Finnasta

WIDE-hackathon
Yazan Alhalabi
Samuel Akangbe
Steven Nebo

Mobiilisovellukset



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

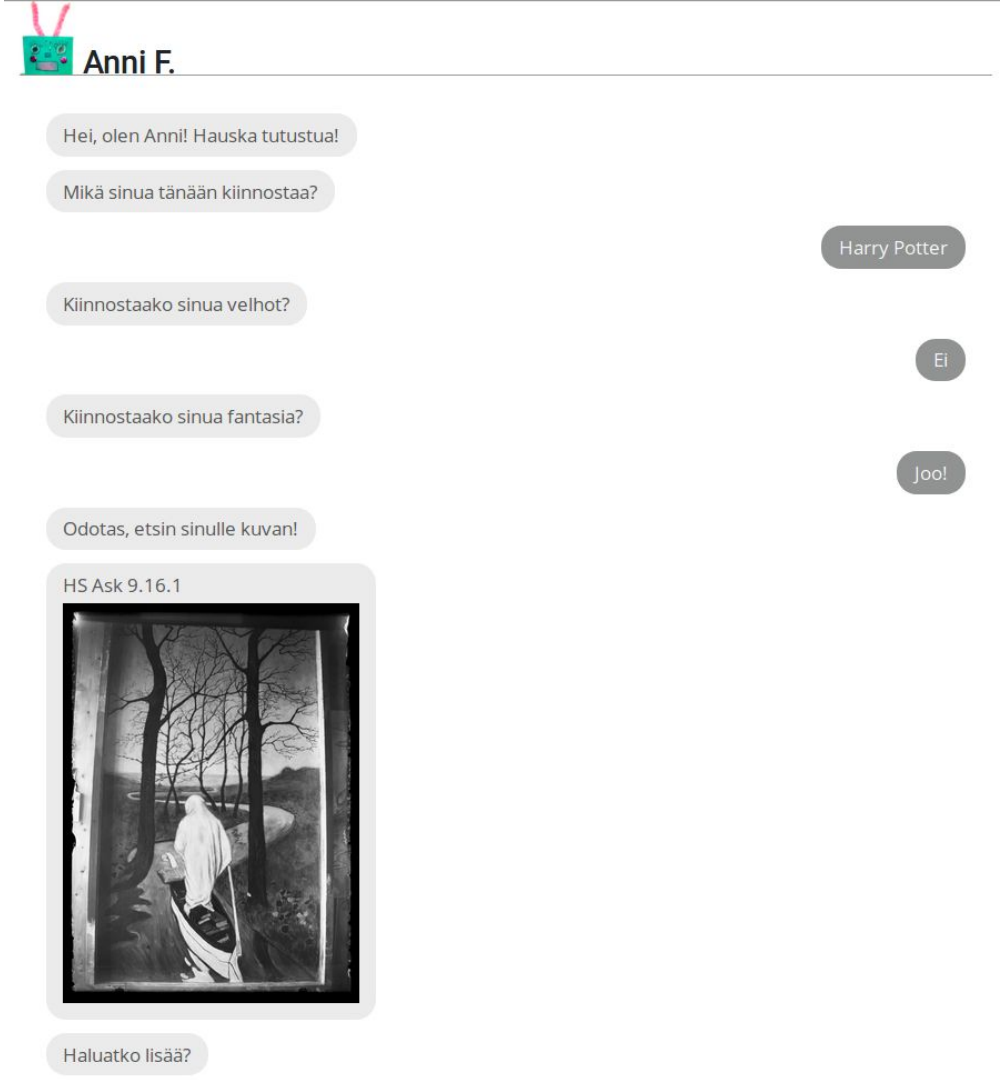


Prototyyppi Android-sovelluksesta, OCR laitteessa
(tekijä Okko Vainonen)

bot.annif.org

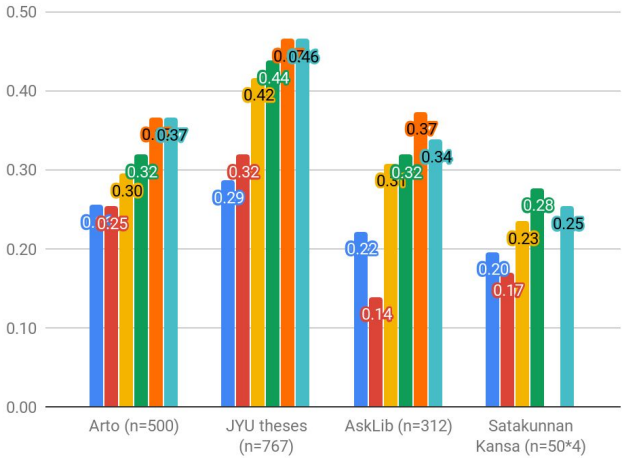
Annifia tekoälynä käyttävä chatbot,
joka suosittelee kuvia ja kirjoja Finnasta

Tervetuloa juttelemaan!



Neljä tavoitetta lähitulevaisuudelle

1. Laadun parantaminen



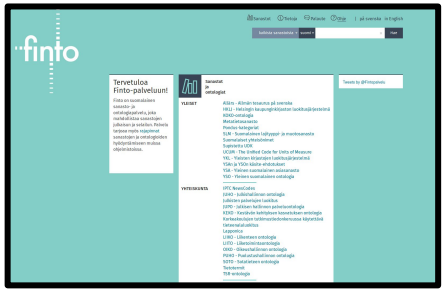
Algoritmien testaus ja arviointi
yhdessä CSC:n kanssa

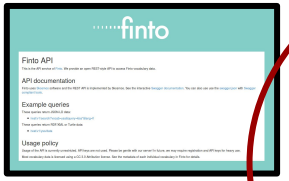
Ihmiset vs. robotit työpaja
Kirjastoverkkopäivillä 2017

KIRJASTOVERKKOPÄIVÄT
Helsinki 23.-24.10.2019
pala palalta parempi

Työpajan nimi	**Työpaja 4** **Automaattinen sisällönkuvailu: Aaveita koneessa**
Kellonaika	9.30–12.00
Paikka	
Yhteyshenkilö	Osma Suominen, Kansalliskirjasto
Työpajan kuvaus	Työpajassa tarkastellaan sisällönkuvailun laatua eri näkökulmista. Käytännön osuudessa arvioidaan ja pisteytetään esimerkkiasiakirjoille tehtyjä sisällönkuvailuja. Tavoitteena on muodostaa yhteinen käsitys siitä, mitä on laadukas sisällönkuvailu sekä vertailla, miten koneellisesti tai koneavusteisesti tuotetut sisällönkuvailut eroavat ihmistyönä tehdyistä. Työpaja on jatkoa vuoden 2017 Kirjastoverkkopäivien Ihmiset vs. robotit -työpajalle.
Kohderyhmä	Sisällönkuvailusta 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

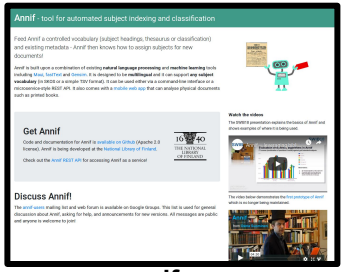




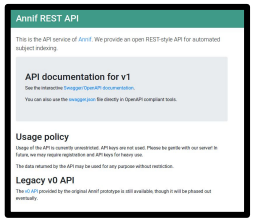
api.finto.fi/rest/
rajapinta sanastodataan
tuotantokäyttöä varten

Uusi Finton automaattisen sisällönkuvailun API

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



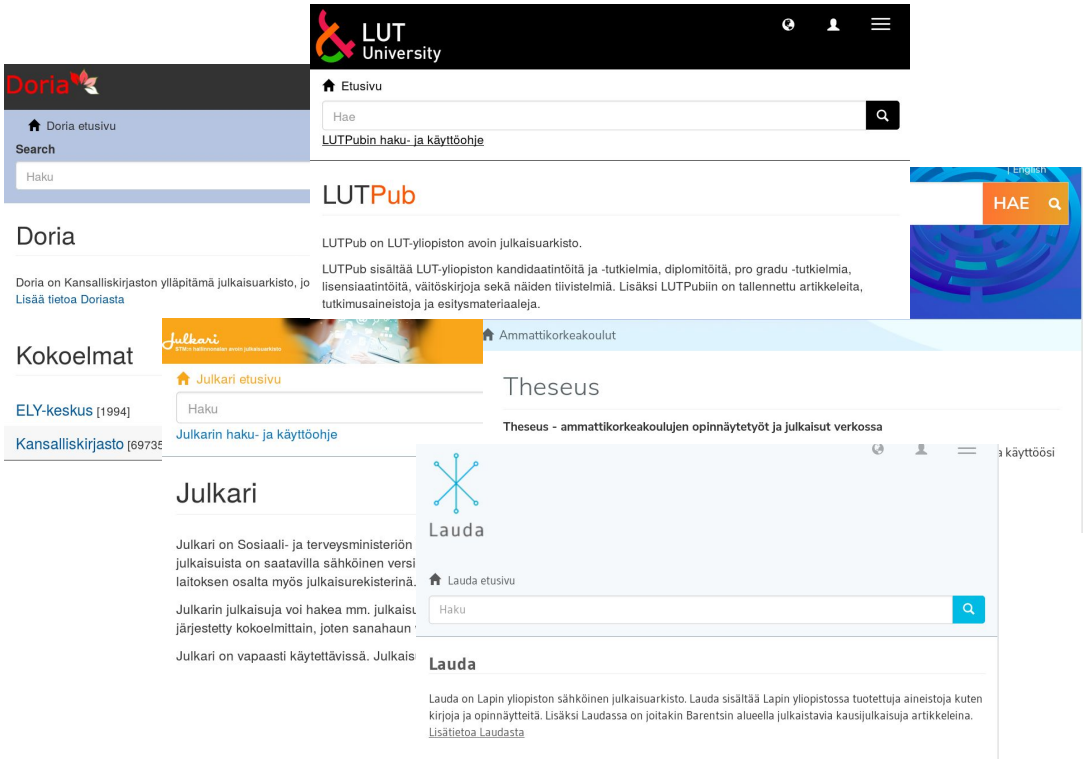
annif.org
Annif-ohjelmiston käyntikorttisivusto



api.annif.org
demorajapinta automaattiseen
sisällönkuvailuuun

Finton rajapinnan laajentaminen automaattisen sisällönkuvailun palveluihin Annifin avulla

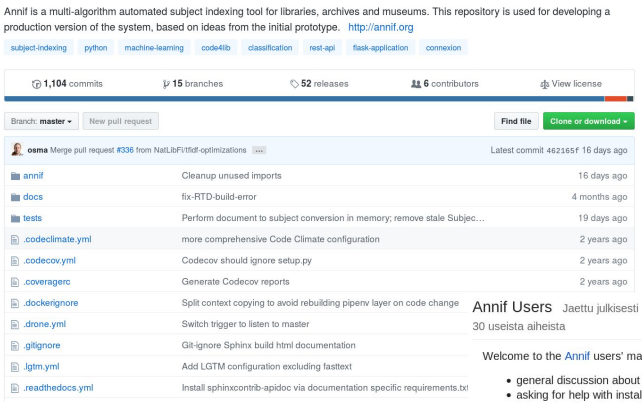
3. Käyttöönotto Kansalliskirjaston järjestelmissä



DSpace-pohjaiset julkaisuarkistot

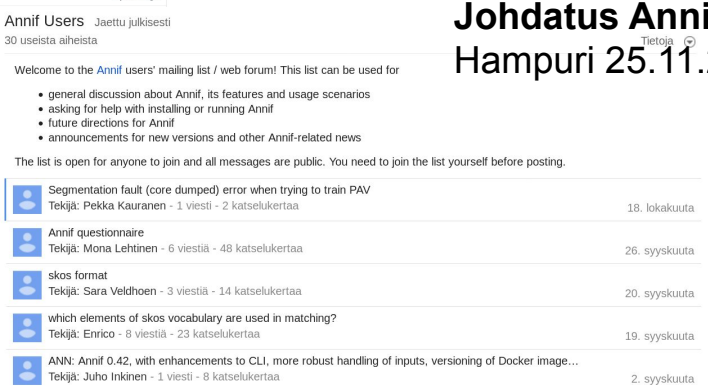
E-vapaakappaleiden vastaanotto ja käsittely

4. Kansainvälinen käyttäjäyhteisö



Koodi, kehitys ja dokumentaatio GitHubissa

Johdatus Annifin käyttöön SWIB19-konferenssissa
Hampuri 25.11.2019



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

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**