



Aalto University
School of Science
Department of Media Technology



Automaattinen asiasanoitus, kieliteknologiset apuvälineet

Eetu Mäkelä

Tiedontuotannon helpottaminen

- Linkitetyn tiedon hyötykäyttö vaatii aineistojen liittämistä käsitteiden URI-tunnisteisiin
valmistusmateriaali: alumiini



valmistusmateriaali: <http://www.yso.fi/onto/koko/p35964>

Tiedontuotannon helpottaminen

- Linkitetyn tiedon hyötykäyttö vaatii aineistojen liittämistä käsitteiden URI-tunnisteisiin
valmistusmateriaali: alumiini



valmistusmateriaali: <http://www.yso.fi/onto/koko/p35964>

- Useissa aineistoissa ei kuitenkaan ole rakenteista metatietoa, vaan ainoastaan tekstikuvauksia
 - [\(Uutis-\)artikkelit](#)
 - [Tapahtumatiedot](#)

Tiedontuotannon helpottaminen

- Linkitetyn tiedon hyötykäyttö vaatii aineistojen liittämistä käsitteiden URI-tunnisteisiin
valmistusmateriaali: alumiini



valmistusmateriaali: <http://www.yso.fi/onto/koko/p35964>

- Useissa aineistoissa ei kuitenkaan ole rakenteista metatietoa, vaan ainoastaan tekstikuvauksia
 - [\(Uutis-\)artikkelit](#)
 - [Tapahtumatiedot](#)
- Tarvitaan palvelu joka pystyy täsmäämään luonnollisen kielen tekstit käsitteisiin

Käsitehaku rekistereistä tekstin perusteella

- “Erkki Tuomiojan mukaan Suomen ulkoministeriön tietomurtoa käsiteltiin tasavallan presidentin Sauli Niinistön kanssa heti tämän lennettyä Helsinkiin.”
- Rekistereissä sanat yleensä perusmuodossa:
 - presidentin -> Wikipedia: [presidentti](#)
- Paitsi jotkut sanat joissakin rekistereissä:
 - lentää -> Wikipedia/KOKO: [lentäminen](#) / [lentäminen](#)
 - presidentti -> KOKO: [presidentit](#)

-> Lexical Analysis Services

- <http://demo.seco.tkk.fi/las/>
- Paketoi mm. [Omorfi-](#), [HFST-](#), [Giellatekno-](#) , [Snowball](#)- ja [langdetect](#)-kieliteknologiaprojektien open source-tuotoksia helposti käytettäväksi web-palveluksi
- Tarjoaa seuraavat palvelut REST/JSON-rajapintoina useille suomalaisesta näkökulmasta kiinnostaville kielille
 - [kielentunnistus](#)
 - [tekstin perusmuotoistus](#)
 - [sanojen taivutus](#)
 - [sanojen muoto-opillinen \(ja semanttinen\) analyysi](#)
 - [sanojen tavutus](#)

-> Automaattinen asiasanoitus: ARPA

Firefox ▾ SPARQL ARPA +

SPARQL ARPA Kieli: suomi | englanti

Teksti:

Erkki Tuomiojan mukaan Suomen ulkoministeriön tietomurtoa käsiteltiin tasavallan presidentin Sauli Niinistön kanssa Helsingissä.

seco.tkk.fi/ssaha/project/resource.shtml?model=dbpedia-fi&uri=http://fi.dbpedia.org/resource/Suomen_ulkoasiainministerio

Erkki Tuomioja mukaan [Suomen ulkoministeriö](#) tietomurto käsittelemisen tasavallan presidentti Sauli Niinistö kanssa [Helsinki](#)

Esimerkkikysely:

DBPedian Suomen laitoksen sisältö ▾

Tekstin käsittely

demo.seco.tkk.fi/ssaha/project/resource.shtml?model=dbpedia-f...ri=http://fi.dbpedia.org/resource/Suomen_ulkoasiainministerio