Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Biografinen datapilvi, muunnokset ja linkitykset

Petri Leskinen
Semantic Computing Research Group (SeCo), Aalto University, http://seco.cs.aalto.fi
petri.leskinen@aalto.fi
Minna Tamper
HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, http://heldig.fi
Semantic Computing Research Group (SeCo), Aalto University, http://seco.cs.aalto.fi
minna.tamper@aalto.fi

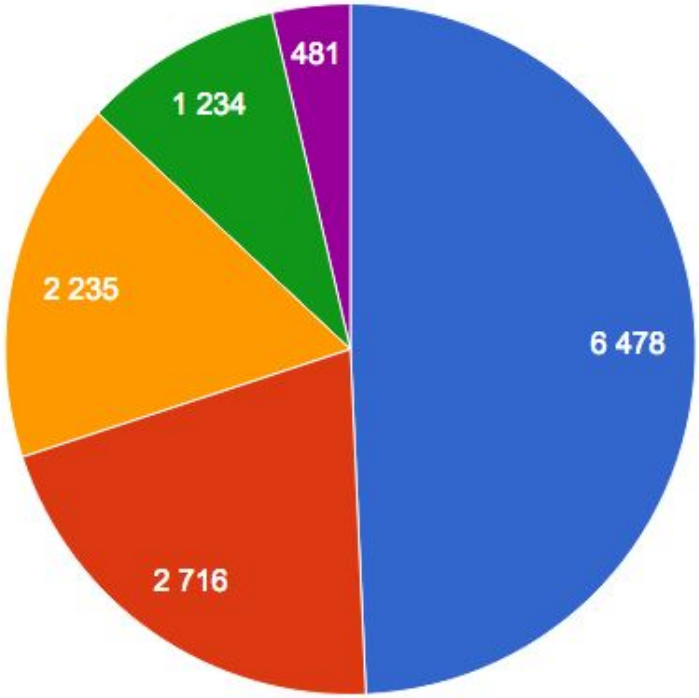
Biografinen datapilvi

- **Alkuperäinen taulukkodata:**
  - Henkilöiden tietokentät: nimi, elinaika, biografiatekstit ...
  - ⇒ Muunnos RDF-muotoon
- **Tekstistä louhitut tietokentät**
  - Henkilöiden sukulaissuhteet, ammatit, tapahtumat
- **Linkitys ulkopuoleisiin tietokantoihin**
  - Yhteensä 14 tietokantaa kuten
    Wikipedia, Fennica, Sotasampo, Geni.com ...
- **Kieliteknologia**
  - Biografioiden käsittely

Biografioita yhteensä 13144

Viisi tietokantaa:

- Kansallisbiografia
- Turun hiippakunnan paimenmuisto 1554–1721
- Talouselämän vaikuttajat
- Suomen papisto 1800–1920
- Kenraalit ja amiraalit



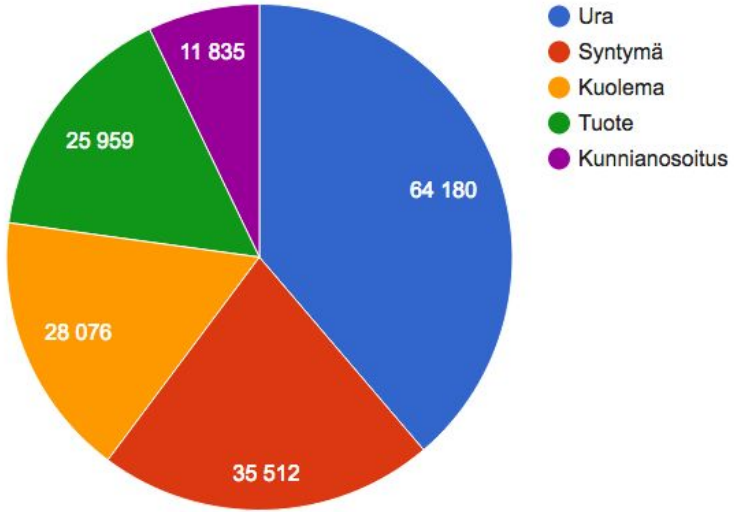
Lisätty henkilödata

Tekstikentistä poimitut sukulaiset ja artikkelien kirjoittajat:

Gottlieb Eliel Saarinen S 20.8.1873 Rantasalmi, K 1.7.1950 Bloomfield Hills, Michigan, Yhdysvallat. V rovasti Juho Saarinen ja Selma Maria Broms. P1 1898–1902 (ero) Mathilda Tony Charlotta Gyldén (sittemmin Gesellius) S 1877, K 1921, P1 V agronomi Axel Gyldén ja Antonia Sofia Hausen; P2 1904– kuvanveistäjä Minna Carolina Louise (Loja) Gesellius S 1879, K 1968, P2 V liikemies Herman Otto Gesellius ja Emilie Struckmann. Lapset: Eva-Lisa (Pipsan) S 1905, K 1979, sisustussuunnittelija, P arkkitehti Jons Robert Ferdinand Swanson; Eero S 1910, K 1961, arkkitehti.

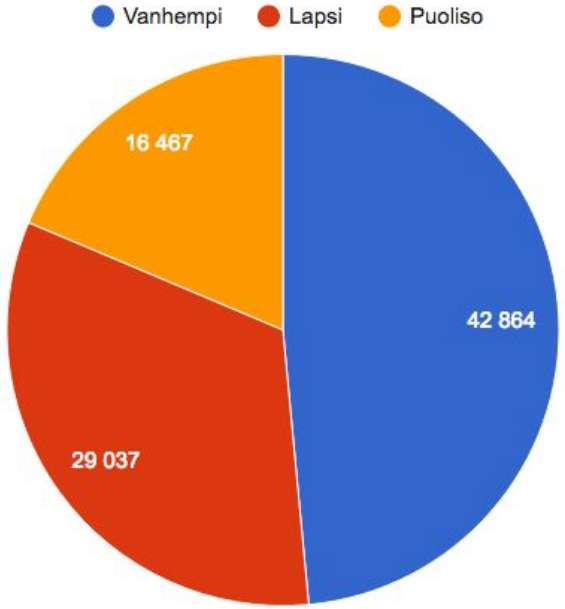
Tekstikentistä poimitut tapahtumat:

Arkkitehtitoimisto Gesellius, Lindgren, Saarinen: Tallbergin talo. 1896–1898, Luotsikatu 1, Helsinki; Pariisin maailmannäyttelyn 1900 paviljonki. 1898–1900, Pariisi; Vakuutusyhtiö Pohjolan talo. 1899–1901, Mikonkatu 3, Helsinki; Pohjoismaiden Osakepankki. 1903–1904, Unioninkatu 32, Helsinki (purettu 1934); Suomen Kansallismuseo. 1902–1911, Helsinki; Helsingin Työväenyhdistyksen talo. 1904, Paasivuorenkatu 5, Helsinki;

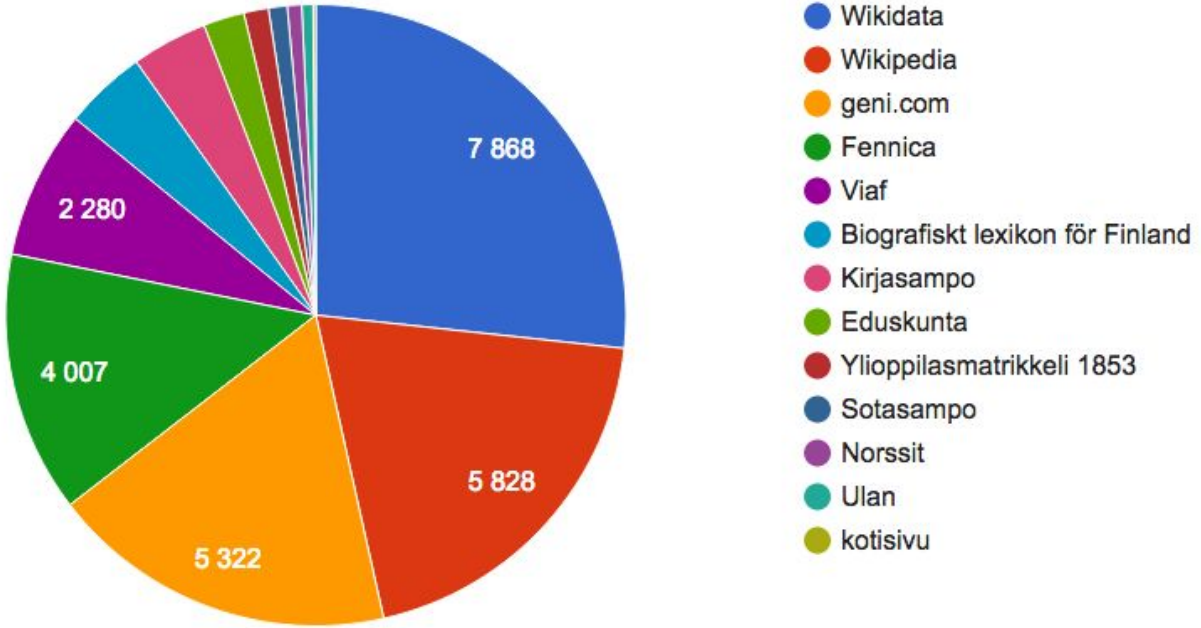


Louhitut sukulaissuhteet

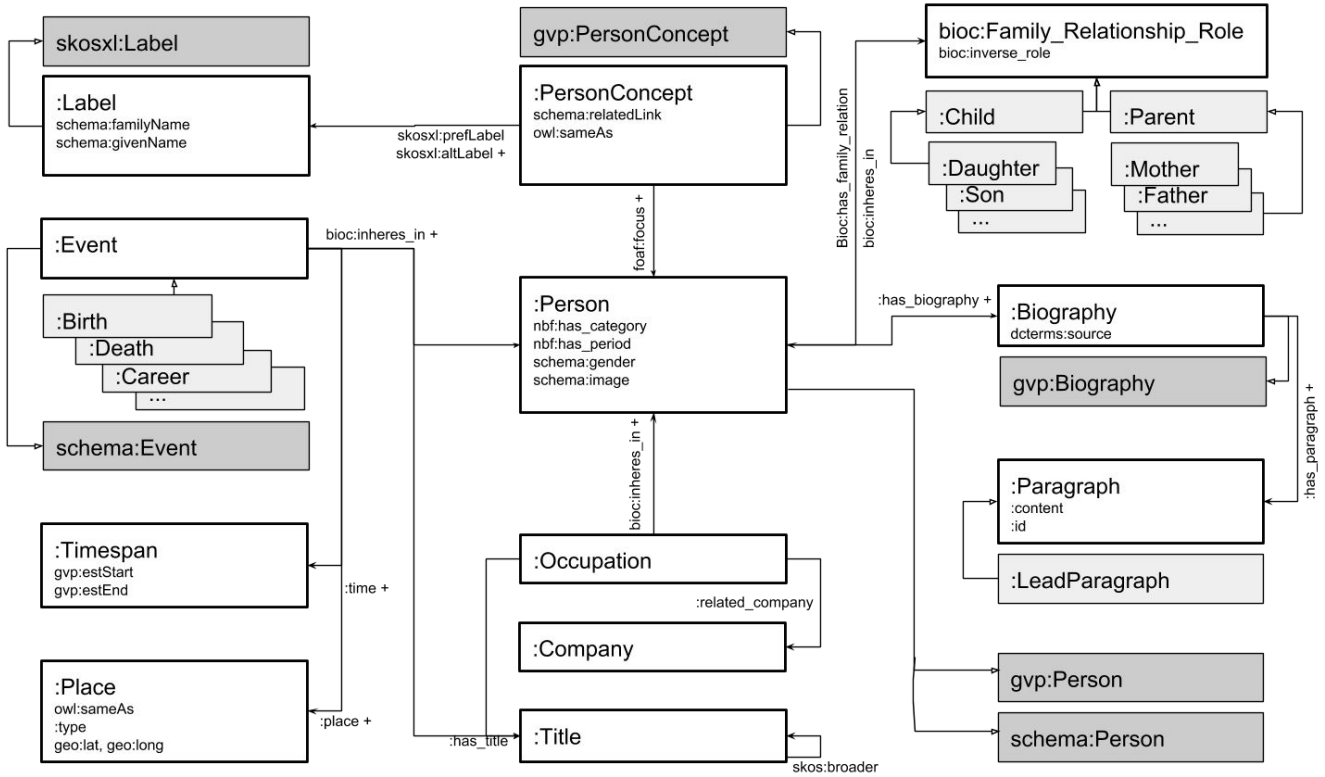
Gottlieb Eliel Saarinen S 20.8.1873 Rantasalmi, K 1.7.1950 Bloomfield Hills, Michigan, Yhdysvallat. V rovasti Juho Saarinen ja Selma Maria Broms. P1 1898–1902 (ero) Mathilda Tony Charlotta Gyldén (sittemmin Gesellius) S 1877, K 1921, P1 V agronomi Axel Gyldén ja Antonia Sofia Hausen; P2 1904– kuvanveistäjä Minna Carolina Louise (Loja) Gesellius S 1879, K 1968, P2 V liikemies Herman Otto Gesellius ja Emilie Struckmann. Lapset: Eva-Lisa (Pipsan) S 1905, K 1979, sisustussuunnittelija, P arkkitehti Jons Robert Ferdinand Swanson; Eero S 1910, K 1961, arkkitehti.



Linkitys ulkopuoleisiin tietokantoihin



Biografiasampo: RDF-datamalli



prefix : <http://ldf.fi/nbf/>

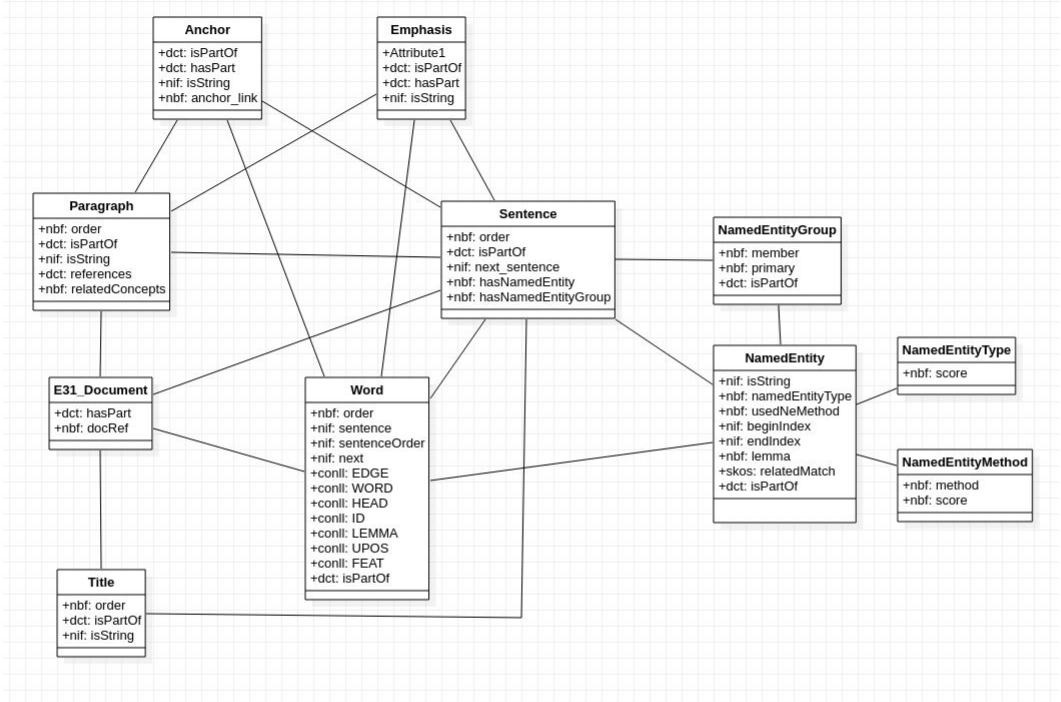
Kieliteknologiset menetelmät

- **Teksti pilkottiin otsikoihin, kappaleisiin, lauseisiin, sanoihin**
- **Tekstissä olevat linkit, korostukset otettiin talteen**
- **Biografioiden annotointi**
  - Finnish-dep-parser, FiNER, CoNLL-RDF, ARPA
  - Vaivannäköä ja koodia
- **Tulokset tallennettiin RDF-formaattiin**

Kieliteknologiolla luotu data

- **n. 114,000,000 tripleä (Elämäkertarakenne)**
- **n. 5 300 000 tripleä (Nimetyt entiteetit)**
- **n. 170 000 tripleä (Tilasto)**
- **n. 10 000 tripleä (Asiasanat)**
- **Linkitys**
  - Biografiasampo (henkilöt, paikat, tittelit)
  - KOKO-ontologia (yhteisöt, organisaatiot)
  - Sotasampo (paikat)

Datamalli


Anchor
+dct: isPartOf
+dct: hasPart
+nif: isString
+nbf: anchor_link
Emphasis
+Attribute1
+dct: isPartOf
+dct: hasPart
+nif: isString
Paragraph
+nbf: order
+dct: isPartOf
+nif: isString
+dct: references
+nbf: relatedConcepts
Sentence
+nbf: order
+dct: isPartOf
+nif: next_sentence
+nbf: hasNamedEntity
+nbf: hasNamedEntityGroup
NamedEntityGroup
+nbf: member
+nbf: primary
+dct: isPartOf
NamedEntity
+nif: isString
+nbf: namedEntityType
+nbf: usedNeMethod
+nif: beginIndex
+nif: endIndex
+nbf: lemma
+skos: relatedMatch
+dct: isPartOf
NamedEntityType
+nbf: score
NamedEntityMethod
+nbf: method
+nbf: score
E31_Document
+dct: hasPart
+nbf: docRef
Word
+nbf: order
+nif: sentence
+nif: sentenceOrder
+nif: next
+conll: EDGE
+conll: WORD
+conll: HEAD
+conll: ID
+conll: LEMMA
+conll: UPOS
+conll: FEAT
+dct: isPartOf
Title
+nbf: order
+dct: isPartOf
+nif: isString