Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Semantic Enrichment

Petri Leskinen & Rafael Leal - 28.11.2024

Data Linking, Reasoning, and Knowledge Extraction

Petri Leskinen

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Overview

- Harmonization
  - Expressions of Time
  - Person names
- Linking
  - External Linking
  - Record Linkage

Harmonization: Expressions of Time

When combining data from multiple sources, several different practises may have been used for indicating the same information

Samples of different data formats with varying precisions:

```
28.4.1884
1884-04-28
1884-1886
1880's
October 1888
Christmas Day 1886
```

Harmonization: Expressions of Time

- Custom Python script for detecting Dates, Years, Seasons, Ranges of Years
- Pattern-based, Regular Expressions
- Two-point or Four-point timespan (CIDOC CRM) responses
  - (start -> end), (start/end of start -> start/end of end)

```
>>> DateConverter.find('10.4.1880')
[[datetime.date(1880, 4, 10), datetime.date(1880, 4, 10), '10.4.1880']]

>>> DateConverter.find('1860, 1862 ja lokakuu 1888')
[[datetime.date(1860, 1, 1), datetime.date(1860, 12, 31), '1860'],
 [datetime.date(1862, 1, 1), datetime.date(1862, 12, 31), '1862'],
 [datetime.date(1888, 10, 1), datetime.date(1888, 10, 31), 'lokakuu 1888']]
```

Person Identification

Same individual can be labeled using multiple name variations

- *Marriage(s)*
- *Finnish/Swedish variations*

Language	Label	Description	Also known as
English	Yrjö Sakari Yrjö-Koskinen	Finnish politician (1830–1903)	Yrjö Koskinen Georg Zacharias Forsman Y. S. Yrjö-Koskinen
Finnish	Yrjö Sakari Yrjö-Koskinen	suomalainen poliitikko	Georg Zacharias Forsman Y. S. Yrjö-Koskinen Yrjö-Sakari Yrjö-Koskinen Yrjö Koskinen Georg Zachris Yrjö-Koskinen Z. Yrjö-Koskinen
Swedish	Yrjö Sakari Yrjö-Koskinen	finsk historiker, senator, professor och publicist	

Parsing Name Variations

- Different practices for writing names
  - Custom pattern based Python scripts for extracting given and family names

```
('Paul Henrik Edelheim',
   [[('Edelheim', 'Paul Henrik')]]),
('Edelheim, Paul Henrik; o. Emilia f. af Brunér', # o. = and, f. = birth name
  [[('Edelheim', 'Paul Henrik')], # person 1
   [('Edelheim (af Brunér)', 'Emilia'), # person 2
      ('Edelheim', 'Emilia'),
      ('af', 'Brunér', 'Emilia')]]),
```

For English see e.g.: https://pypi.org/project/probablepeople/

Entity Linking

- ARPA-UI Service
- MediaWiki API Query Service

Entity Linking

- ARPA-UI Service
  - Search by entity name

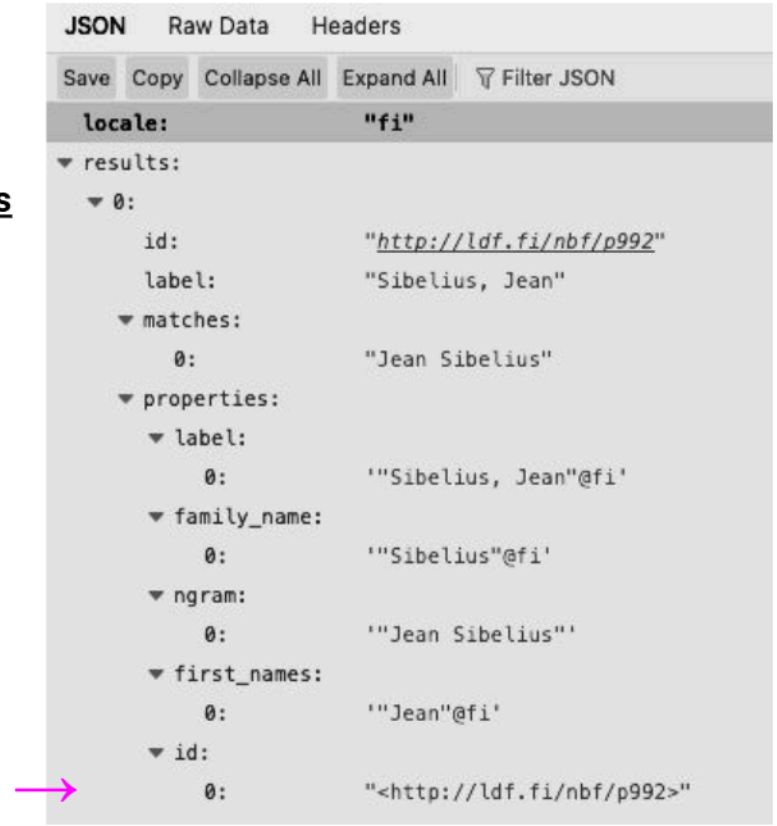
    http://demo.seco.tkk.fi/arpa/nbf_person?text=Jean+Sibelius

    http://demo.seco.tkk.fi/arpa/wikidata-organizations?text=Stockmann

Entity Linking

Search for *"Jean Sibelius"* in BiographySampo publication

http://demo.seco.tkk.fi/arpa/nbf_person?text=Jean+Sibelius



External Linking

- MediaWiki API Query Service
  - Example search by person name https://api.triplydb.com/s/LhiRugIR4

```
VALUES (?search_string) {
  ("Aino Järnefelt")
  ("Emil Kivi")
  ("Rudyard Kipling")
}
BIND(wd:Q5 AS ?human_class)

SERVICE wikibase:mwapi {
    bd:serviceParam wikibase:endpoint "www.wikidata.org";
                    wikibase:api "EntitySearch";
                    mwapi:search ?search_string ;
                    mwapi:language "fi" .
    ?wikidata_ID wikibase:apiOutputItem mwapi:item .
    ?score wikibase:apiOrdinal true .
}
```

External Linking

- MediaWiki API Query Service
  - Example search by person name https://api.triplydb.com/s/LhiRugIR4

search_string	wikidata_ID	wikiLabel	score	gender	birth_time	death_time	biographysampo_ID
Aino Järnefelt	wd:Q406078	Aino Sibelius	0	female	1871-08-10T00:00:00Z	1969-06-08T00:00:00Z	
Emil Kivi	wd:Q17381689	Kaarlo Eemeli Kivirikko	0	male	1870-01-28T00:00:00Z	1947-04-26T00:00:00Z	p4347
Rudyard Kipling	wd:Q34743	Rudyard Kipling	0	male	1865-12-30T00:00:00Z	1936-01-18T00:00:00Z	
Rudyard Kipling	wd:Q120482345	Rudyard Kipling Sullivan	31		1906-01-01T00:00:00Z	1977-01-01T00:00:00Z	

Record Linkage and Deduplication

p3325168332085551640	Forsman, G. Z.	G. Z.		Forsman	male		1879	1879	1879
p2332799835176310591	Forsman, Gabriel	Gabriel		Forsman	male		1870	1870	1870
p1644249866634343660	Forsman, Georg Jakob	Georg Jakob		Forsman	male		1846	1849	1852
p2566107394316456339	Forsman, Georg Zachris	Georg Zachris		Forsman	male		1867	1867	1867
p2452958869818531794	Forsman, Gustaf Erik	Gustaf Erik		Forsman	male		1853	1853	1853

? ↕ matching between several datasets

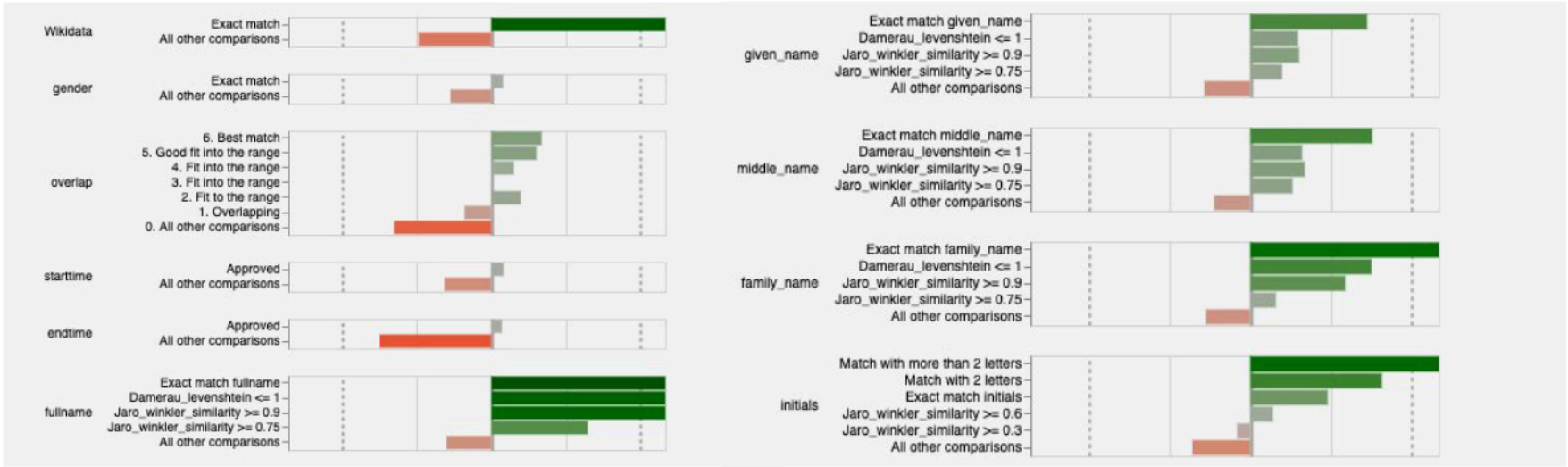
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Georg Zachris	Georg Zachris	Yrjö-Koskinen	male	http://www.wikidata.org/entity/Q3771268	Wikidata		1	1830-12-10	Vaasa	h	1	1	1903-11-13	Helsinki				
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Yrjö-Sakari	Yrjö-Sakari	Yrjö-Koskinen	male	http://www.wikidata.org/entity/Q3771268	Wikidata		1	1830-12-10	Vaasa	h	1	1	1903-11-13	Helsinki				
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Yrjö Sakari	Yrjö Sakari	Yrjö-Koskinen	male	http://www.wikidata.org/entity/Q3771268	Wikidata		1	1830-12-10	Vaasa	h	1	1	1903-11-13	Helsinki				
http://ldf.fi/yoma/people/p16509	Koskinen, Yrjö	Yrjö	Koskinen	male	http://www.wikidata.org/entity/Q3771268	Wikidata		1	1830-12-10	Vaasa	h	1	1	1903-11-13	Helsinki				
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Yrjö Sakari	Yrjö Sakari	Yrjö-Koskinen	male	http://edelfelt.sls.fi/personer/1723/yrjo-koskir	Edelfelt		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Forsman, Georg Zacharias	Georg Zacharias	Forsman	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Georg Zakarias	Georg Zakarias	Yrjö-Koskinen	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Koskinen, Yrjö Sakari	Yrjö Sakari	Koskinen	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Forsman, Yrjö Sakari	Yrjö Sakari	Forsman	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Yrjö Sakari	Yrjö Sakari	Yrjö-Koskinen	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Koskinen, Yrjö	Yrjö	Koskinen	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Z.	Z.	Yrjö-Koskinen	male	http://urn.fi/URN:NBN:fi:au:finaf:000063708	KANTO		1	1830-12-31			1	1	1903-12-31					
http://ldf.fi/yoma/people/p16509	Yrjö-Koskinen, Yrjö Sakari	Yrjö Sakari	Yrjö-Koskinen	male	http://www.yso.fi/onto/kaunokki#person_123	BookSampo		1	1830-12-10	Vaasa	h	1	1	1903-11-13	Helsinki				
http://ldf.fi/yoma/people/p16509	Forsman, Georg Zacharias	Georg Zacharias	Forsman	male	http://www.yso.fi/onto/kaunokki#person_123	BookSampo		1	1830-12-10	Vaasa	h	1	1	1903-11-13	Helsinki				

Record Linkage and Deduplication

- **Python Record Linkage Toolkit**
  - Customizable for string distances
  - For small datasets
- **Dedupe**
  - Unsupervised learning
  - Term Frequency learning
- **SPLINK**
  - Active learning

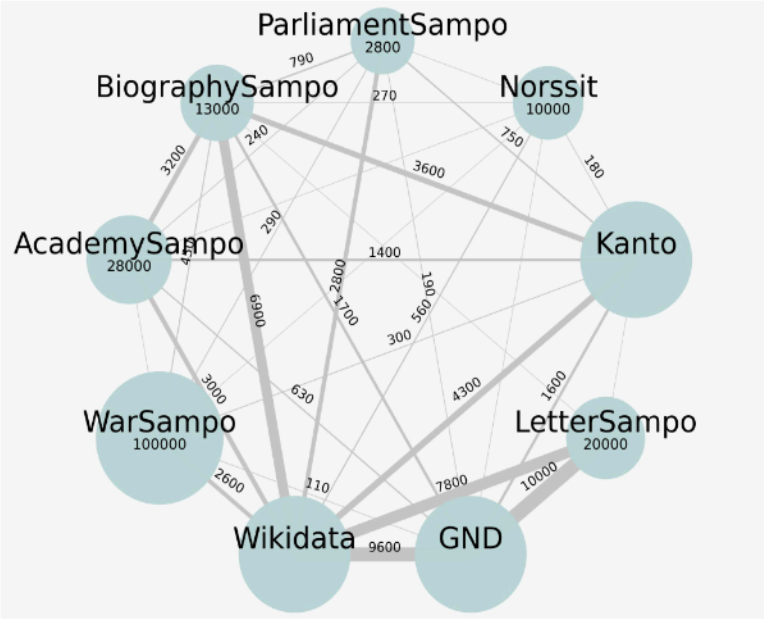
Record Linkage and Deduplication

Analyze the results of SPLINK



Upcoming project ”SampoSampo”

- Combine data from Sampo projects and external LOD sources
- People, Places, and Organizations
  - *~200.000 historical Finnish people*
  - *~50.000 organizations*
  - *~20.000 geographical locations*



NLP within the Sampo framework

Rafael Leal

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Metadata enrichment

- Makes the data easier to handle (computers) and explore (end-users)
- Enables new kinds of analysis

Examples of existing tools

- **Keywords**
- **Named entities**

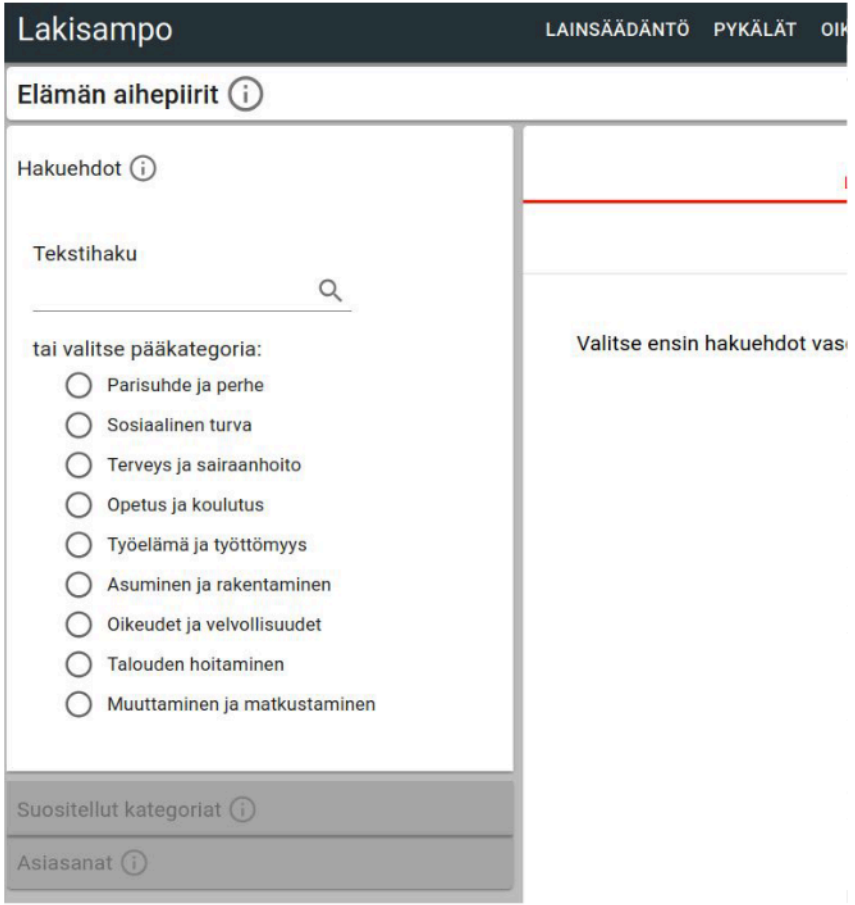
Keyword extraction

- **Annif**
  - Tool for automated subject indexing and classification
  - Developed by the National Library of Finland
  - Returns YSO concepts as keywords
  - Used for example by Yle to assign tags to news articles

Keyword extraction

Topics

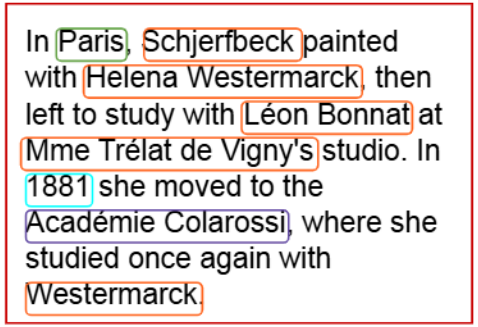
LawSampo



Rafael Leal et al.: **Relevance Feedback Search Based on Automatic Annotation and Classification of Texts**. 2021.

Named entity recognition and linking

- **Entity extraction**
- **Sometimes hard to define**
- **Finnish tools**
  - TurkuNLP Group's NER tool
    - *FinBERT*
  - Stanza



Named entity recognition and linking

- **Candidate identification**
  - Vector-based
    - *Search for similar vectors in a vector space*
    - *Better recall*
  - Lexical
    - *Search for similar word forms from a list*
    - *Better precision*

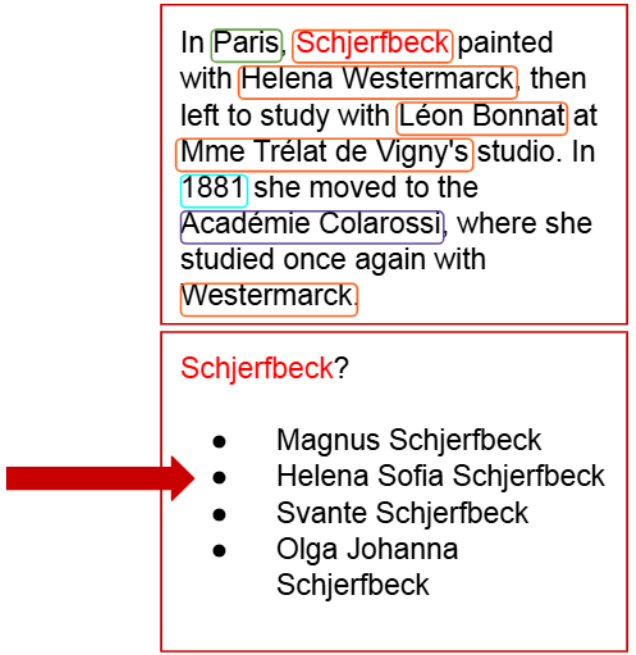
In Paris, Schjerfbeck painted with Helena Westermarck, then left to study with Léon Bonnat at Mme Trélat de Vigny's studio. In 1881 she moved to the Académie Colarossi, where she studied once again with Westermarck.

Schjerfbeck?

- Magnus Schjerfbeck
- Helena Sofia Schjerfbeck
- Svante Schjerfbeck
- Olga Johanna Schjerfbeck

Named entity recognition and linking

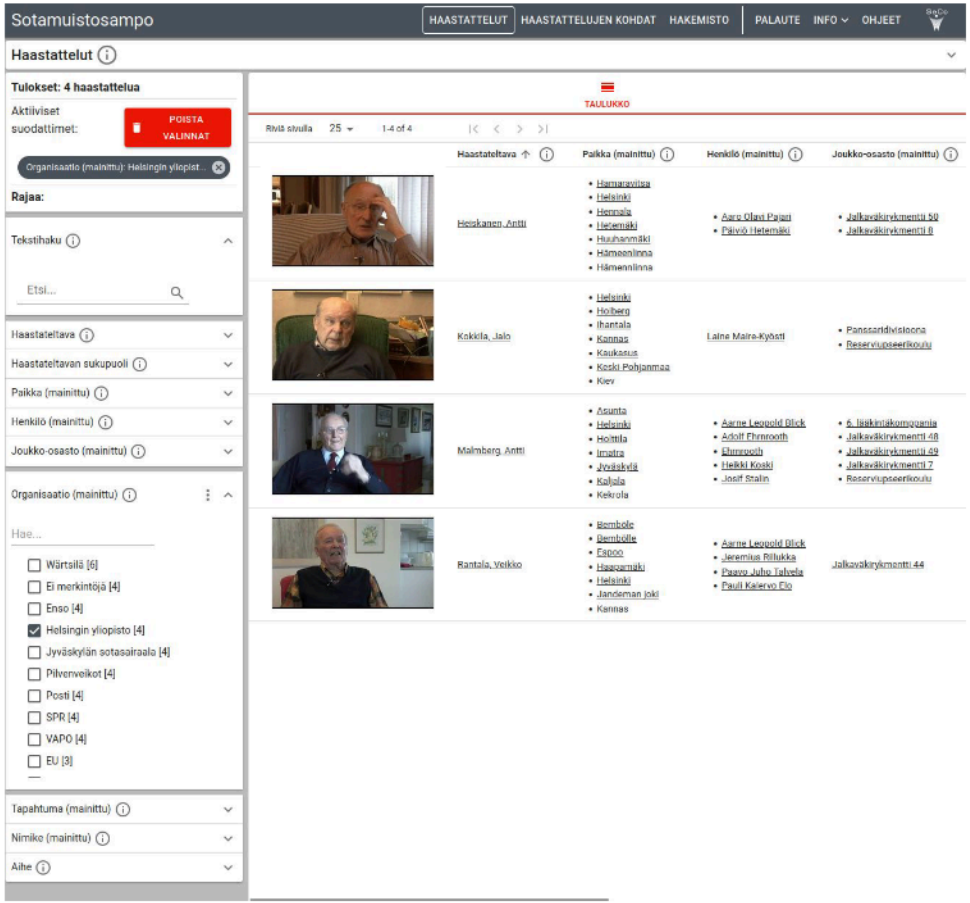
- **Disambiguation and Linking**
  - Manual
  - Graph-based
  - Vector-based
    - *LLMs*



Named entity recognition and linking

Disambiguation and Linking

WarMemoirSampo



Mikko Koho et al.: **Building Lightweight Ontologies for Faceted Search with Named Entity Recognition: Case WarMemoirSampo**. *2022*.

Aalto University
School of Science

UNIVERSITY OF HELSINKI

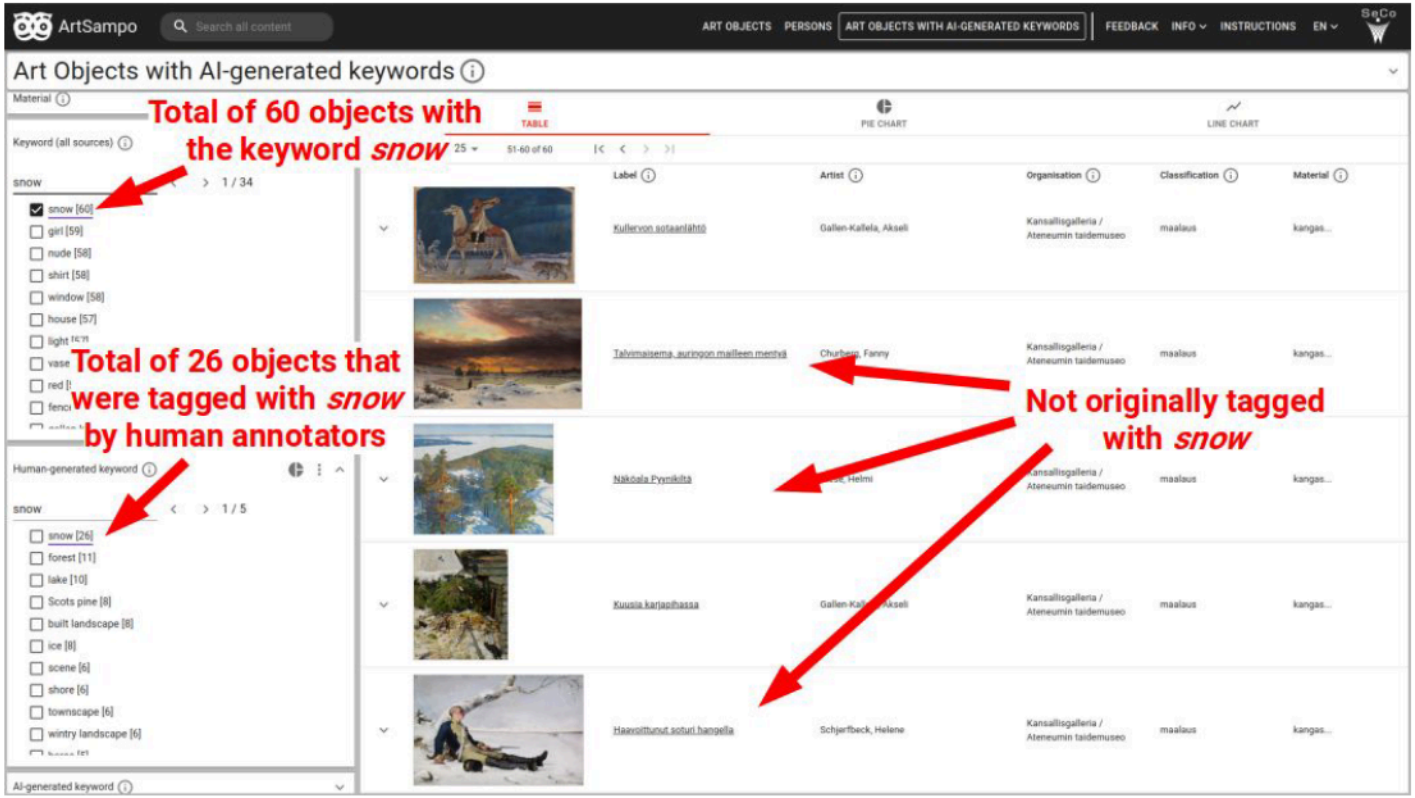
SeCo

HELDIG
Helsinki Centre for Digital Humanities

Using LLMs in NLP

- **Usually no need to fine-tune**
  - Or fine-tuning via examples in the prompt
- **Do not blindly trust the contents!**
  - Next word prediction → hallucinations
- **Reasoning**
  - RAG (Retrieval-Augmented Generation)

ArtSampo

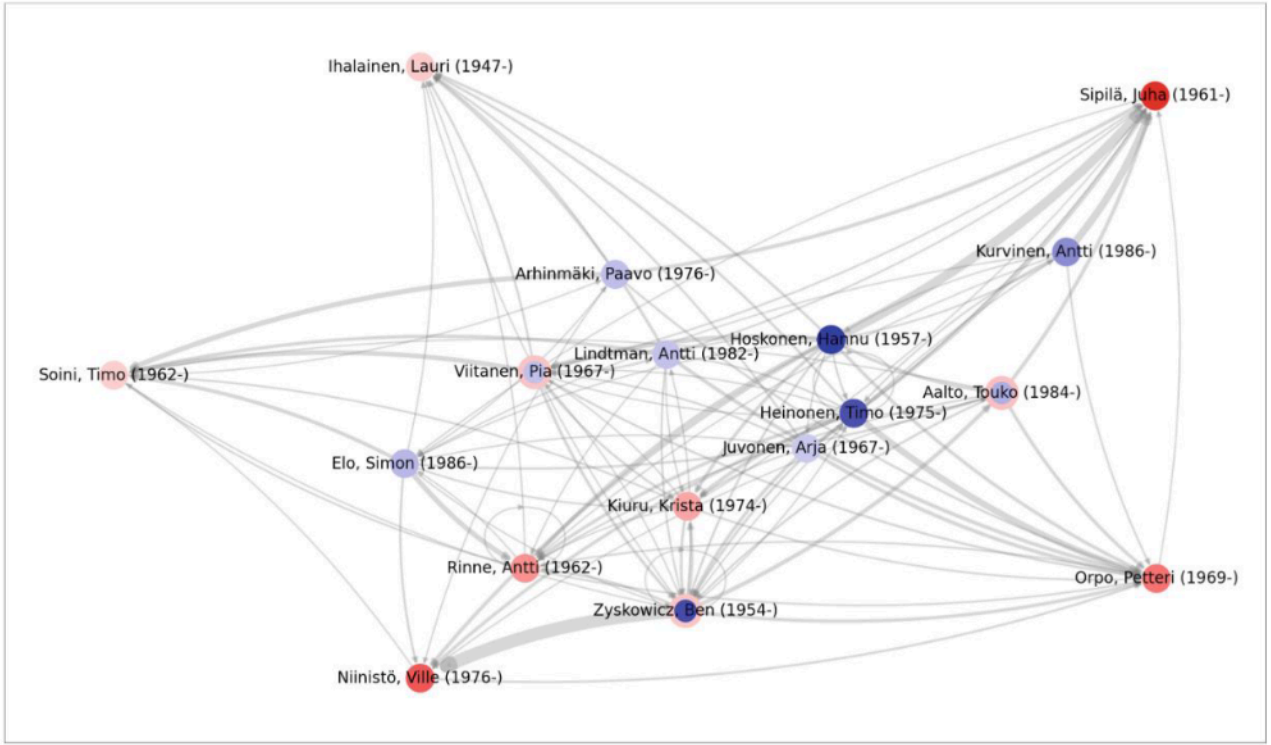


Annastiina Ahola et al.: **Using generative AI and LLMs to enrich art collection metadata for searching, browsing, and studying art history in Digital Humanities**. 2024.

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix sne: <http://ldf.fi/snellman/entities/> .
@prefix snl: <http://ldf.fi/snellman/> .
@prefix sns: <http://ldf.fi/snellman/schema/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

snl:3865 sns:entity_instance sne:i1e419f4422a59ef_7ebed50db3e6f7d,
        sne:i5357f195eeccbcc_ec78efefee298c5,
        sne:i624d9d29224f519_b1b3dbdc0dc9718,

sne:e1e419f4422a59ef a sns:Entity ;
        sns:entity_category <Place> ;
        sns:instance sne:i1e419f4422a59ef_7ebed50db3e6f7d ;
        owl:sameAs snl:13397,
        <http://www.yso.fi/onto/yso/p94148> ;
        skos:prefLabel "Hämeenlinna" .

sne:e5357f195eeccbcc a sns:Entity ;
        sns:entity_category <Person> ;
        dcterms:description "1824–1909. Snellmanin velipuoli. Maanviljelijä
Alahärmässä." ;
        sns:instance sne:i5357f195eeccbcc_ec78efefee298c5 ;
        owl:sameAs snl:12459 ;
        skos:prefLabel "Henrik Albert Snellman" .

sne:i1e419f4422a59ef_7ebed50db3e6f7d a sns:EntityInstance ;
        sns:end_char "1419"^^xsd:Integer ;
        sns:start_char "1405"^^xsd:Integer ;
        sns:surface_form "Hämeenlinnassa" .

sne:i5357f195eeccbcc_ec78efefee298c5 a sns:EntityInstance ;
        sns:end_char "1052"^^xsd:Integer ;
        sns:start_char "1046"^^xsd:Integer ;
        sns:surface_form "Albert" .
```

ParliamentSampo

Ten MPs with highest hub and authority values



Henna Poikkimäki et al.: **Analyses of Networks of Politicians Based on Linked Data: Case ParliamentSampo - Parliament of Finland on the Semantic Web**. 2022.

References

Annastiina Ahola, Lilli Peura, Rafael Leal, Heikki Rantala and Eero Hyvönen: **Using generative AI and LLMs to enrich art collection metadata for searching, browsing, and studying art history in Digital Humanities**. *Proceedings, 2nd International Conference on Data & Digital Humanities Generative Artificial Intelligence for Text and Multimodal Data 12th - 13th December 2024, University of Minho, Braga, Portugal*, November, 2024. Accepted, forth-coming.

Mikko Koho, Rafael Leal, Esko Ikkala, Minna Tamper, Heikki Rantala and Eero Hyvönen: **Building Lightweight Ontologies for Faceted Search with Named Entity Recognition: Case WarMemoirSampo**. *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022)* (Sanju Tiwari, Nandana Mihindukulasooriya, Francesco Osborne, Dimitris Kontokostas, Jennifer D'Souza and Mayank Kejriwal (eds.)), vol. 3184, pp. 19-35, CEUR Workshop Proceedings, May, 2022. International Knowledge Graph Generation From Text (TEXT2KG).

Rafael Leal, Joonas Kesäniemi, Mikko Koho and Eero Hyvönen: **Relevance Feedback Search Based on Automatic Annotation and Classification of Texts**. *3rd Conference on Language, Data and Knowledge (LDK 2021)* (Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo and Barbara Heinisch (eds.)), Open Access Series in Informatics (OASIcs), vol. 93, pp. 18:1-18:15, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

Henna Poikkimäki, Petri Leskinen, Minna Tamper and Eero Hyvönen: **Analyses of Networks of Politicians Based on Linked Data: Case ParliamentSampo - Parliament of Finland on the Semantic Web**. *New Trends in Database and Information Systems*, pp. 585-592, Springer International Publishing, August, 2022.