

Tools, knowledge extraction, and data production

DARIAH Aalto Day 2022, November 30th

Content

- **Sampo-UI for User Interface Design**
- **Knowledge extraction from texts**
- **Data production pipelines**

Sampo-UI for User Interface Design

Heikki Rantala
Annastiina Ahola

Sampo-UI

- **Framework for developing user interfaces for semantic portals in accordance with Sampo model**
 - multiple application perspectives for same data
 - Combining faceted search with various visualizations
 - two-step usage cycle
- **Built with modern JavaScript libraries and data queried with SPARQL**
- **Goal: Making it easier to create highly customizable and responsive user interfaces for semantic portals**
 - Configurable through portal and perspective specific JSON files
 - Comes with ready-to-use data-analytic tooling that can be easily expanded based on a portal's needs

Building Apps with Sampo-UI

- A working application can be built in only a couple of hours using the existing elements
- Doesn't require a lot of expertise in JavaScript programming
- New “modules” (visualizations etc) can be added to the framework and then reused in other applications
- <https://github.com/SemanticComputing/sampo-ui>

Sampo-UI in action

Multiple perspectives for viewing the data

The screenshot displays the Sampo-UI interface. At the top, there is a navigation bar with the Sampo-UI logo, a search bar, and links to PERSPECTIVE 1, PERSPECTIVE 2, PERSPECTIVE 3, CLIENTFS, FEEDBACK, INFO, INSTRUCTIONS, and EN. The main header features a background image of old books with the text "Sampo-UI" and the quote "Here to forge for us the Sampo, Hammer us the lid in colors". Below the header, a prompt says "Select a perspective to search and browse the knowledge graph:". Four perspective cards are shown: Perspective 1 (manuscript), Perspective 2 (open book), Perspective 3 (gavel), and ClientFS (world map). Two orange arrows point from the text "Multiple perspectives for viewing the data" to the Perspective 1 and Perspective 3 cards. At the bottom, there is a footer with logos for Aalto University School of Science, UNIVERSITY OF HELSINKI, HELDIG, and SeCo.

Sampo-UI

Search all content

PERSPECTIVE 1 PERSPECTIVE 2 PERSPECTIVE 3 CLIENTFS FEEDBACK INFO INSTRUCTIONS EN

Sampo-UI

"Here to forge for us the Sampo,
Hammer us the lid in colors"

Select a perspective to search and browse the knowledge graph:

Perspective 1
Perspective 1 description

Perspective 2
Perspective 2 description

Perspective 3
Perspective 3 description

ClientFS
Client-side faceted search

Images used under license from Shutterstock.com

Aalto University School of Science

UNIVERSITY OF HELSINKI

HELDIG
Helsinki Centre for Digital Humanities

Sampo-UI in action

Faceted search options

Search all content

Perspective 1 **Perspective 2** PERSPECTIVE 3 CLIENTFS FEEDBACK INFO INSTRUCTIONS EN

Results: 222605 manuscripts

Narrow down by:

Label

Search...

Author

Search...

- ☐ Augustine, Saint, Bishop of Hippo [4661]
- ☐ Jerome, Saint, 419 or 420 [3245]
- ☐ Cicero, Marcus Tullius [2575]
- ☐ Anonymous [2327]
- ☐ Aristotle [2029]
- ☐ Gregory, I, Pope, approximately 540-604 [1980]
- ☐ Bernard, of Clairvaux, Saint, 1090 or 1091-1153 [1864]
- ☐ John Chrysostom, Saint, 407 [1748]
- ☐ Aquinas, Thomas, Saint, 1225-1274 [1477]
- ☐ Bede, the Venerable, Saint, 673-735 [1420]

Work

Production place

Production date

Note

Language

TABLE PRODUCTION PLACES PRODUCTION HEATMAP PRODUCTION DATES EVENT DATES LAST KNOWN LOCATIONS MIGRATIONS NETWORK EXPORT

Rows per page 10 1-10 of 222605

Label	Author	Work	Expression	Production place	Production date	Last known loc.
Montpellier (F), BU Historique de Médecine, H 069	-	-	-	-	-	-
Montpellier (F), BU Historique de Médecine, H 073	-	-	-	-	-	-
Paris (F), Bibliothèque nationale de France, Manuscrits, Coll. Dupuy 653	-	-	-	-	-	-
Paris (F), Bibliothèque nationale de France, Manuscrits, lat. 04950	Trogus, Pompeius	Historiarum epitome libris quadraginta quatuor absoluta: praemittuntur omnium librorum prologi	Historiarum epitome libris quadraginta quatuor absoluta: praemittuntur omnium librorum prologi	-	0800 - 0901 0801-01-01 - 0900-12	Paris
Montpellier (F), BU Historique de Médecine, H 075	-	-	-	-	1201-01-01 - 1300-12	-
Montpellier (F), BU Historique de Médecine, H 076	-	-	-	-	0901-01-01 - 1100-12	-
Montpellier (F), BU Historique de Médecine, H 077	-	-	-	-	0801-01-01 - 1000-12	-
Montpellier (F), BU Historique de Médecine, H 078	-	-	-	-	1101-01-01 - 1300-12	-
Montpellier (F), BU Historique de Médecine, H 079	-	-	-	-	1301-01-01 - 1400-12	-

Tabs for different formats of showing the data (table, visualizations..)

Results set

Figure 1 consists of three parts. The top part contains two line graphs showing 'Manuscript production by decade'. The first graph, 'Manuscript production by decade (all)', shows a sharp peak in the 13th century, reaching over 60,000 manuscripts. The second graph, 'Manuscript production by decade (France)', shows a much lower production, with a small peak in the 13th century and a larger one in the 15th century. The bottom part is a network graph showing relationships between manuscripts. Nodes are labeled with manuscript IDs, and edges represent relationships. The graph is dense and complex, with many nodes and edges. A legend on the left indicates the types of relationships: 'MS' (Manuscript), 'MS' (Manuscript), and 'MS' (Manuscript).

Configuring Sampo-UI

```
"portalID": "sampo",
"rootUrl": "",
"perspectives": {
  "searchPerspectives": [
    "perspective1",
    "perspective2",
    "fullTextSearch"
  ],
  "onlyInstancePages": [
    "manuscripts",
    "places"
  ]
},
"localeConfig": {
  "defaultLocale": "en",
  "readTranslationsFromGoogleSheets": false,
  "availableLocales": [
    {
      "id": "en",
      "label": "English",
      "filename": "localeEN.json"
    }
  ]
},
"sitemapConfig": {
  "baseUrl": "https://sampo-ui.demo.seco.cs.aalto.fi",
  "langPrimary": "en",
  "outputDir": "../src/server/sitemap_generator",
  ...
}
```

Settings concerning the whole portal (included perspectives, language settings, endpoint for queries, etc.) are configured in the portal specific configuration JSON file.

Configuring Sampo-UI

```
"id": "perspective2",
  "endpoint": {
    "url": "https://ldf.fi/mmm/sparql",
    "useAuth": false,
    "prefixesFile": "SparqlQueriesPrefixes.js"
  },
  "sparqlQueriesFile": "SparqlQueriesPerspective2.js",
  "baseURI": "http://ldf.fi/mmm",
  "URITemplate":
"<BASE_URI>/manifestation_singleton/<LOCAL_ID>",
  "facetClass": "frbroo:F4_Manifestation_Singleton",
  "frontPageImage": "main_page/manuscripts-452x262.jpg",
  "searchMode": "faceted-search",
  "defaultActiveFacets": [],
  "defaultTab": "table",
  "defaultInstancePageTab": "table",
  "includeInSitemap": true,
  "resultClasses": {
    "perspective2": {
      "paginatedResultsConfig": {
        "tabID": 0,
        "component": "ResultTable",
        "tabPath": "table",
        "tabIcon": "CalendarViewDay",
        "propertiesQueryBlock":
"manuscriptPropertiesFacetResults",
        "paginatedResultsAlwaysExpandRows": false,
        "paginatedResultsRowContentMaxHeight": 190,
        "pagesize": 10
      },
      "instanceConfig": {
        ...
```

Each perspective has its own configuration file for configuring the results view(s) and instance pages for the result set objects.

The facets included in the facet search options and the properties included in the table for the perspective can be independently configured in this configuration file.

Knowledge extraction from texts

Minna Tamper
Rafael Leal
Petri Leskinen

Background

- **Combining NLP and Linked Data**
 - Two-way process
- **Content extraction and enrichment**
 - More expressive and more useful ontologies
 - NER, NEL, search, classification, etc
- **Applied in Sampo Portals**
 - LawSampo, ParliamentSampo, BiographySampo, WarMemoirSampo, among others
 - New facets and applications

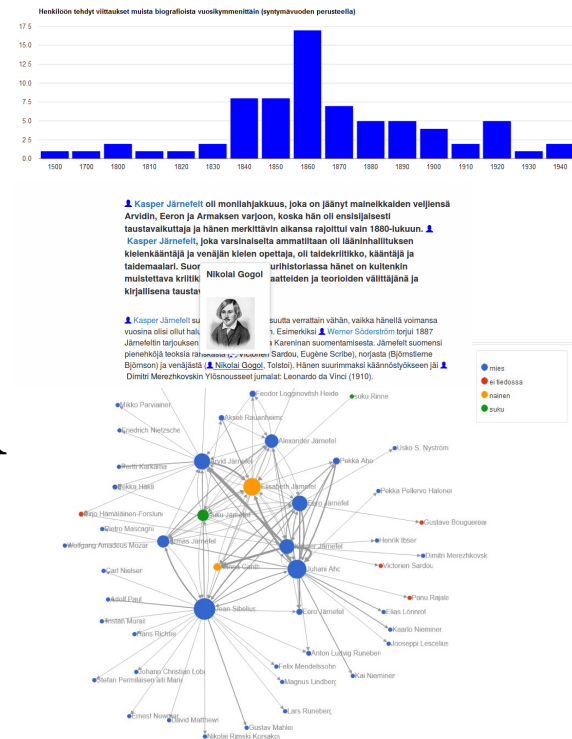
NLP methods and applications

- **Secompling**

- Finnish NLP library
- Integration of in-house and 3rd party tools
 - *FastText, Annif, BERT, TurkuNLP tools*
- Lemmatization, Keyword extraction, Unsupervised classification, Keyword-based search engine, Flexible language identification

NLP methods and applications

- **Named Entity Extraction and Linking pipeline**
 - Identifying and linking named entities in text to given ontology
 - Components: NER (combined Finnish NER tools), Entity Linking, Disambiguation
 - Application: pseudonymization, social network analysis, search application, etc.



Providing NLP Tooling for Linked Data as Web Services

- **Goal: provide documented NLP services as easy to deploy packages (plug and play) for users**
- **To be published as part of the Linked Data Finland platform**
 - <https://nlp.ldf.fi>

Select a tool of choice

FINER

Named entity recognition based on the FIN-CLARIN's FINER tool.

FinBERT

TurkuNLP's named entity recognition application based on Google's BERT.

APPI - Automatic annotation tool

Named entity linking based on voting scheme.

Finnish dependency parser

Linguistic analysis based on the Turku NLP Finnish-dep-parser

Date and registry number identifiers

Regular expression based application for identifying registry numbers, dates, and more.

Person name finder service

Application that can pick up person names from text and offers links and contextual information (gender, lifespan) about each name.

Binary Gender Identification Service

Application that can determine person's gender based on a given name usage statistics.



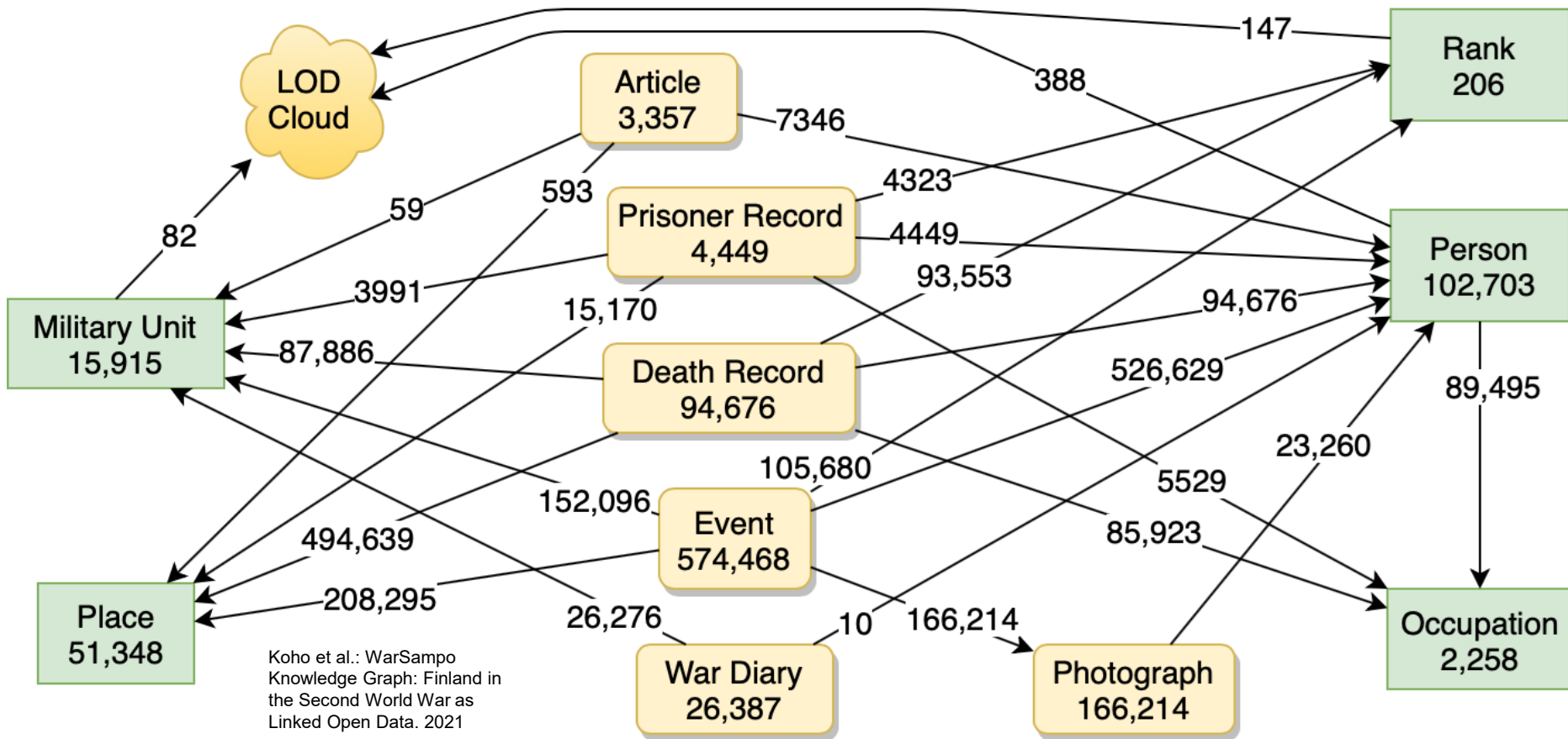
Providing NLP Tooling for Linked Data as Web Services

- **Wrapping NLP services from other research projects**
- **Turku NLP Tools**
 - FinBERT models such as NER, Turku Neural Parser,
- **Aalto NLP tools**
 - Arpa, LAS, Anoppi, Nelli tagger, name-finder, reksi, ...
- **Currently, portal under development**
 - Swagger API documentation
 - Demo UI
 - Collecting and dockerizing new tools

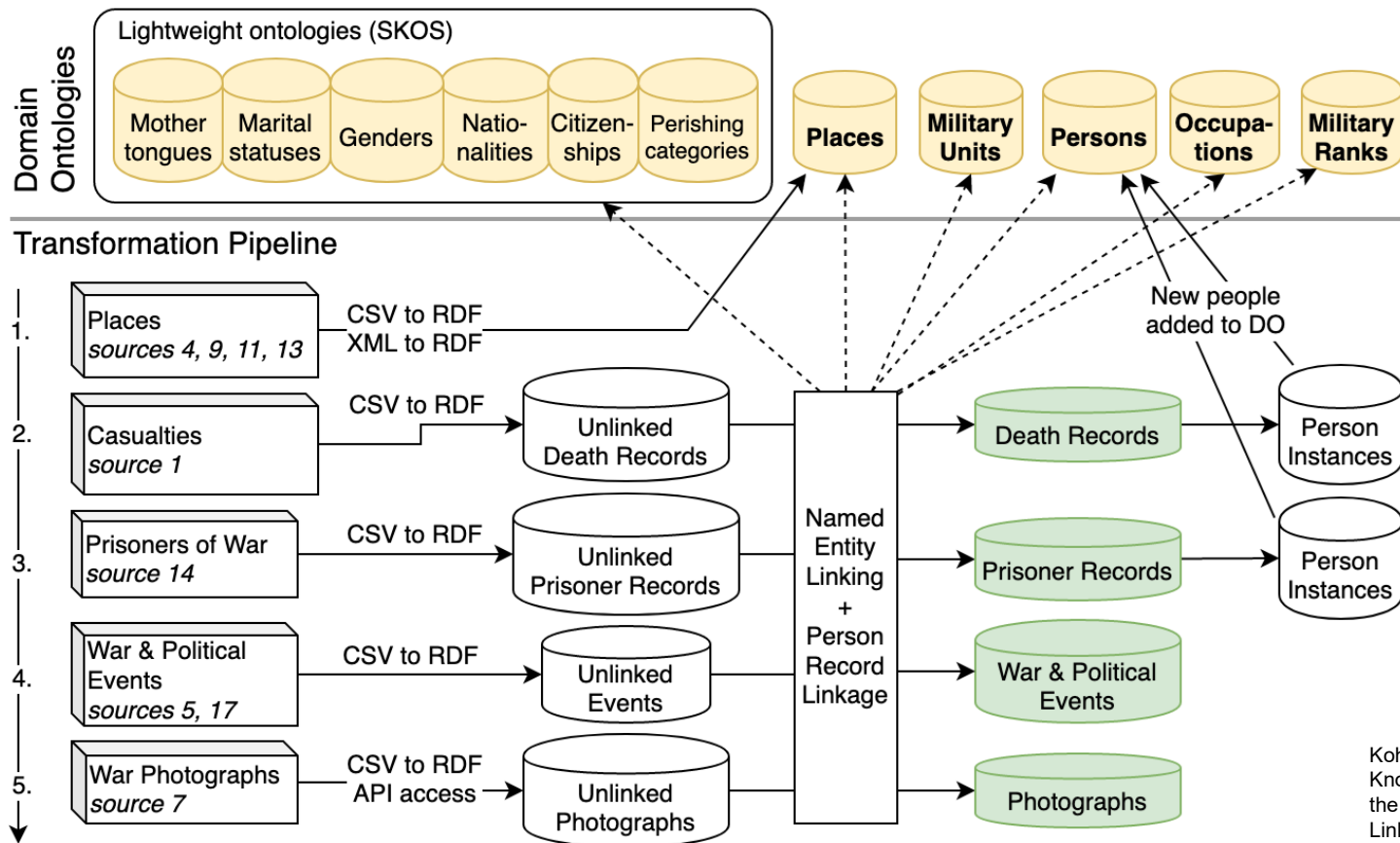
Data Production Pipelines

Mikko Koho

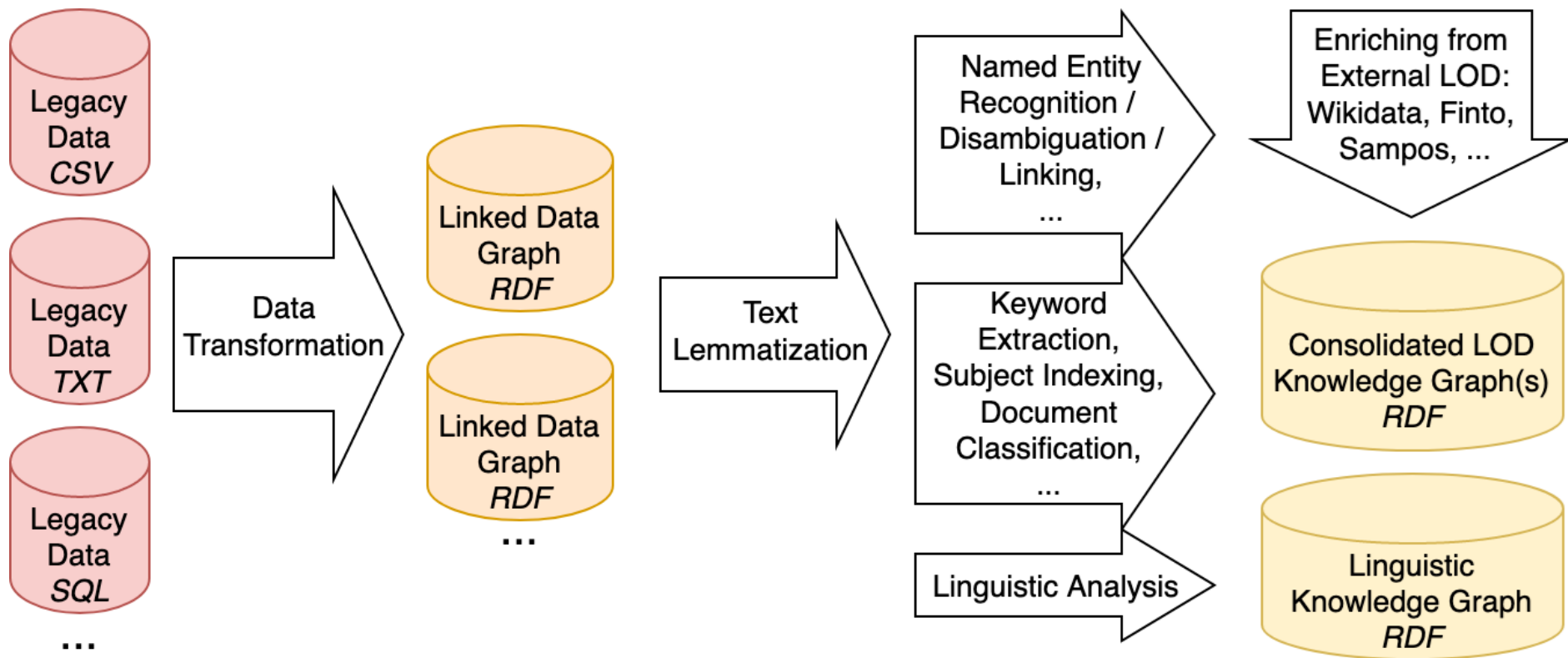
WarSampo: Maintenance challenge



WarSampo Data Transformation Pipeline



Data Production & Maintenance Model



Repeatable Data Production Pipelines

- **Transform heterogeneous source datasets into a consolidated LOD knowledge graph**
- **Interlinking between datasets/entities**
- **Enrich data from external LOD sources**
- **Based on containers and high performance computing**
 - Repeatable: re-run when source data is updated
- **Computationally intensive tasks into one pipeline**
 - Data transformations, NLP tasks, disambiguation, etc.
- **Prototyping manual maintenance with WarSampo KG**

Thank you!



Aalto University
School of Science



UNIVERSITY OF HELSINKI



HELDIG
Helsinki Centre for Digital Humanities