Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

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

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Sampo-UI for User Interface Design

Heikki Rantala
Annastiina Ahola

Sampo-UI

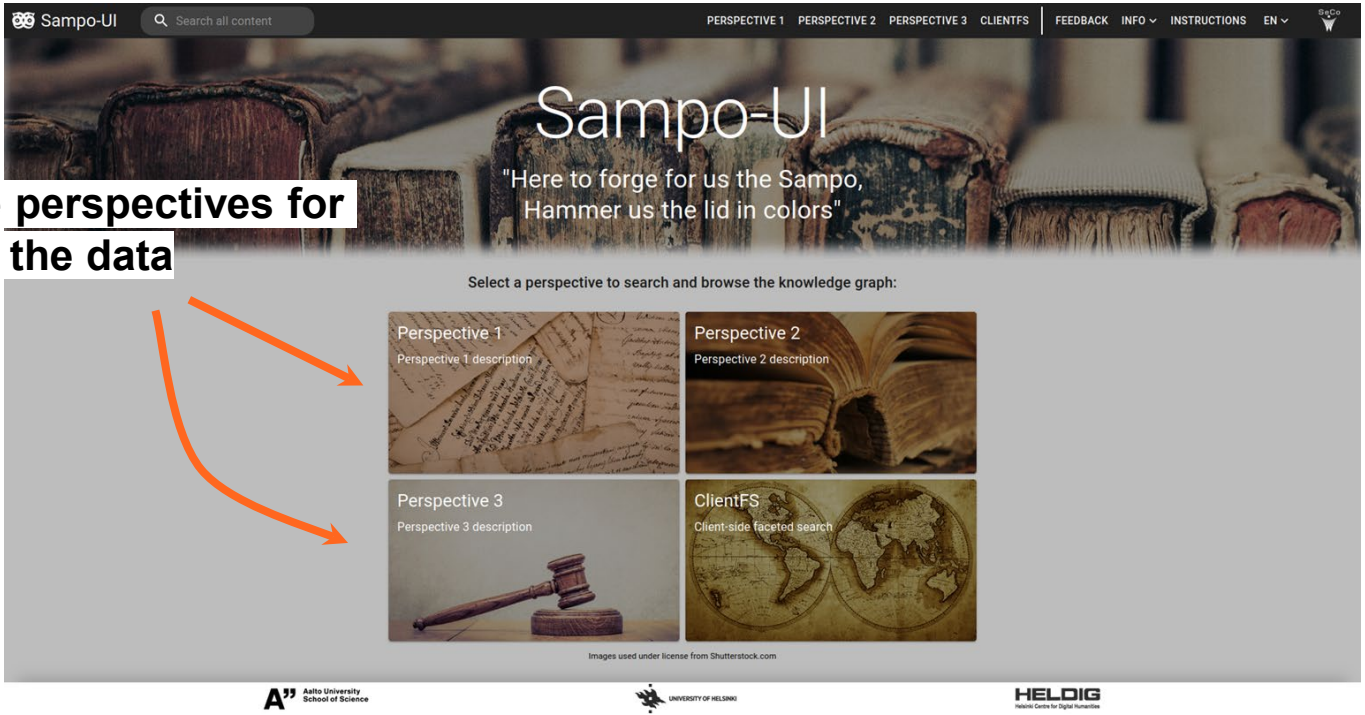
- **Framework for developing user interfaces for semantic portals in accordance with Sampo model**
  - multiple application perspectives for same data
  - Combining faceted search with various visualizations
  - two-step usage cycle
- **Built with modern JavaScript libraries and data queried with SPARQL**
- **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

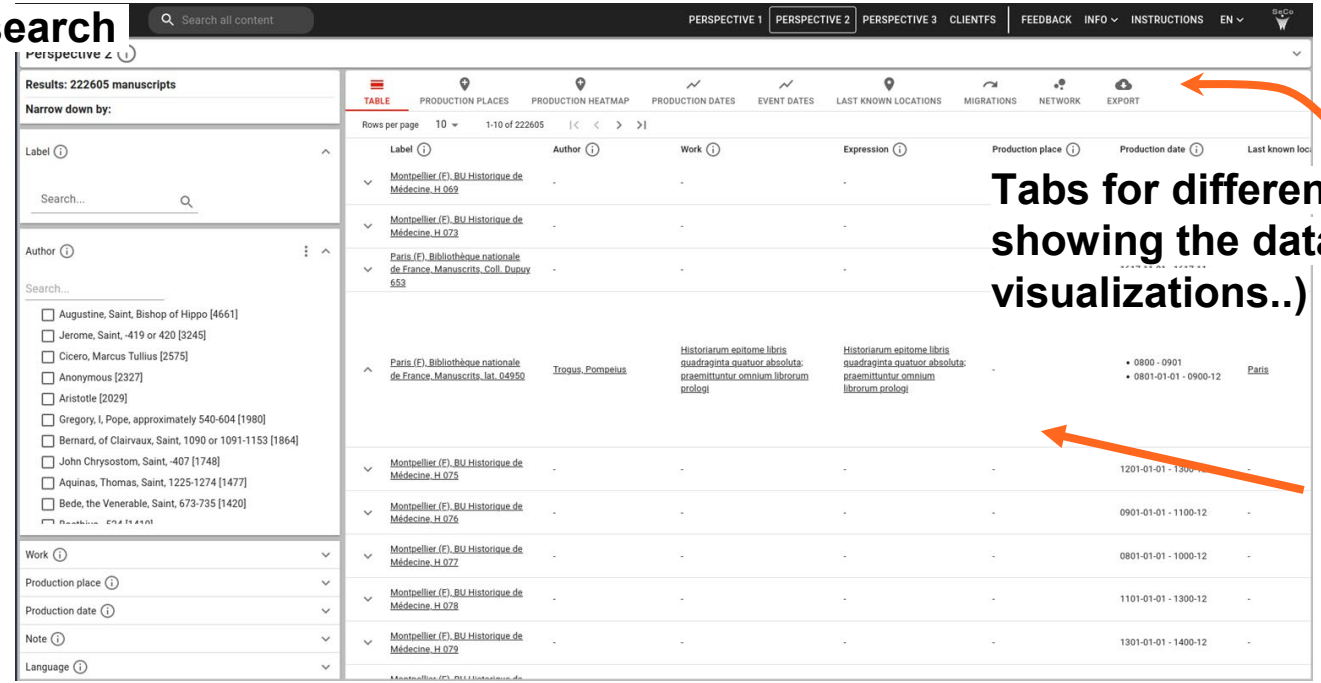
Sampo-UI in action

Multiple perspectives for viewing the data



Sampo-UI in action

Faceted search options

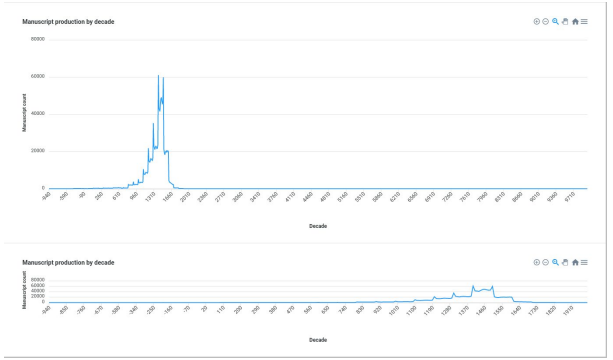


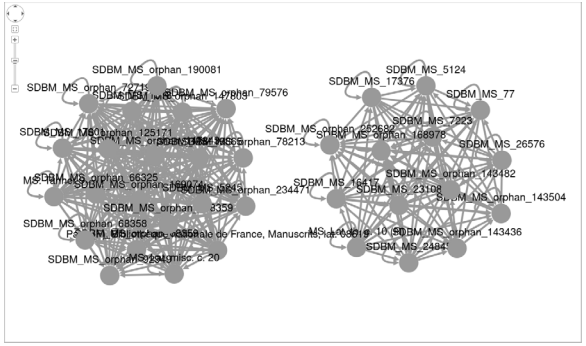
Tabs for different formats of showing the data (table, visualizations..)

Results set

Sampo-UI in action

Ready-to-use components for visualizing result data





Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

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.

Aalto University
School of Science

UNIVERSITY OF HELSINKI

SeCo

HELDIG
Helsinki Centre for Digital Humanities

Knowledge extraction from texts

Minna Tamper
Rafael Leal
Petri Leskinen

Background

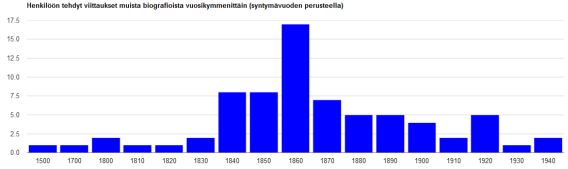
- **Combining NLP and Linked Data**
  - Two-way process
- **Content extraction and enrichment**
  - More expressive and more useful ontologies
  - NER, NEL, search, classification, etc
- **Applied in Sampo Portals**
  - LawSampo, ParliamentSampo, BiographySampo, WarMemoirSampo, among others
  - New facets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

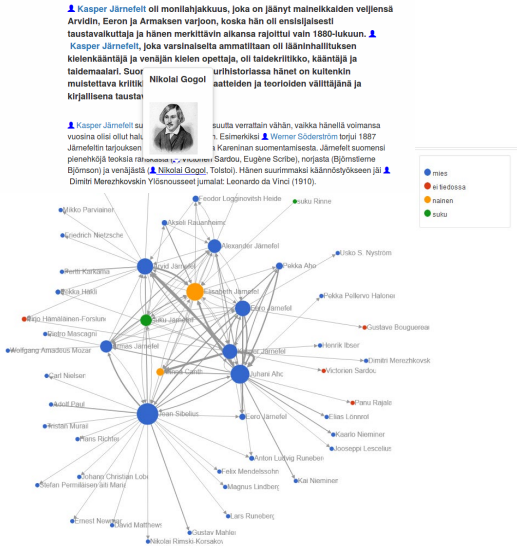
NLP methods and applications

- **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

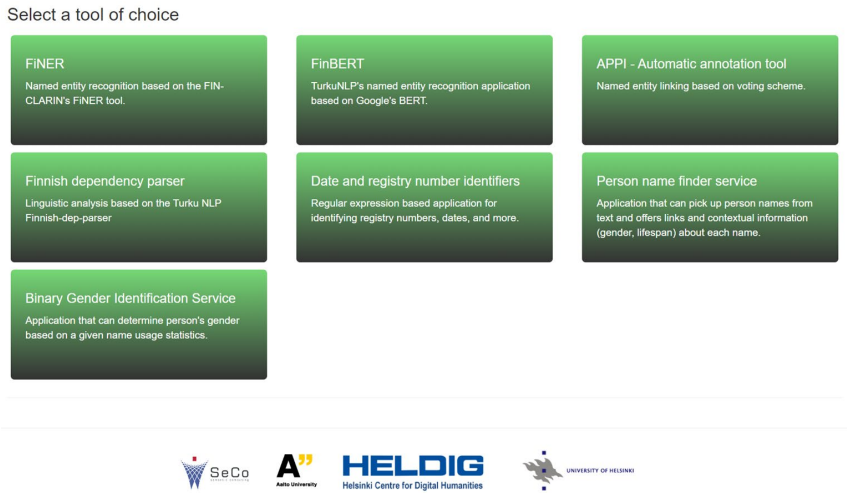
- **Named Entity Extraction and Linking pipeline**
  - Identifying and linking named entities in text to given ontology
  - Components: NER (combined Finnish NER tools), Entity Linking, Disambiguation
  - Application: pseudonymization, social network analysis, search application, etc.





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



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

Aalto University
School of Science

UNIVERSITY OF HELSINKI

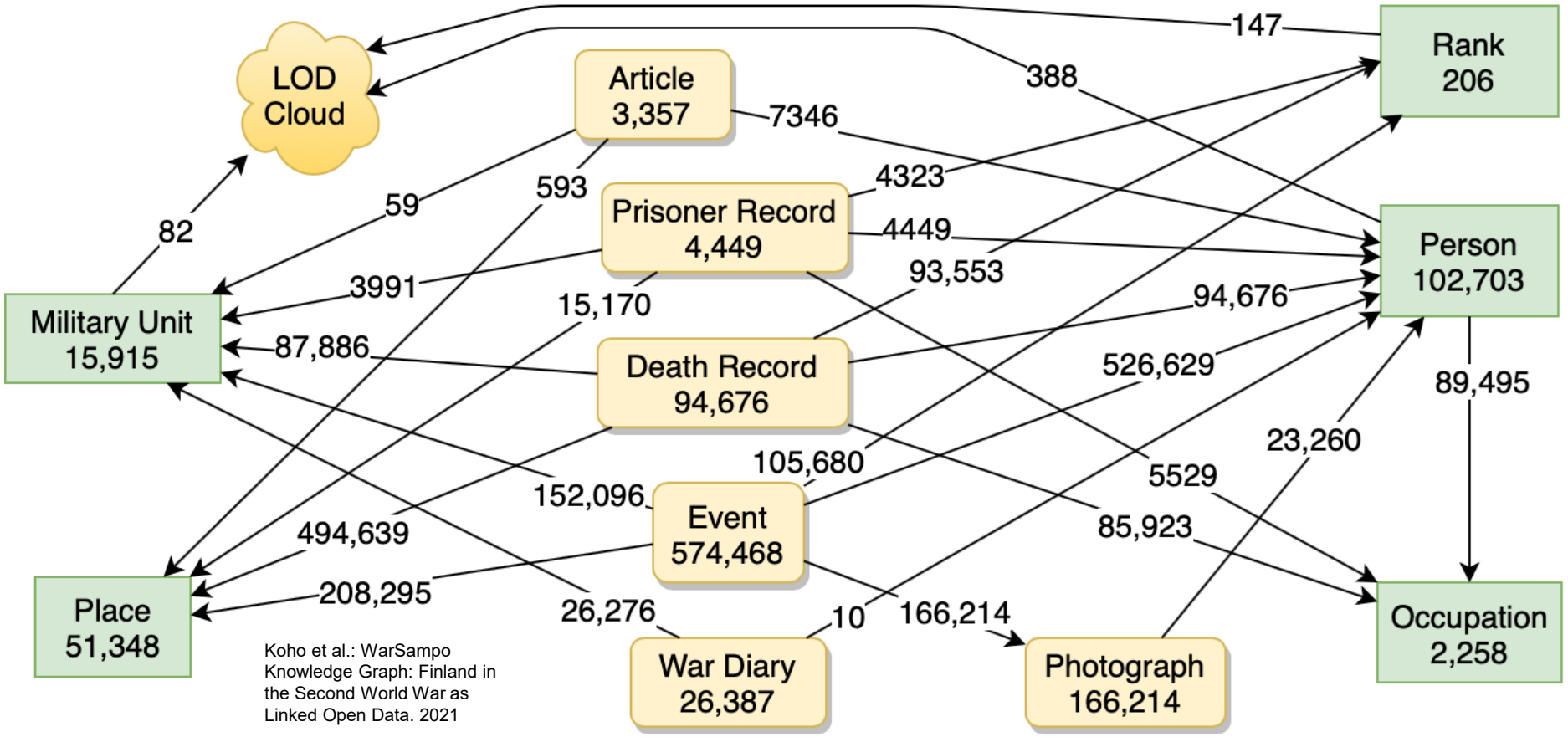
SeCo

HELDIG
Helsinki Centre for Digital Humanities

Data Production Pipelines

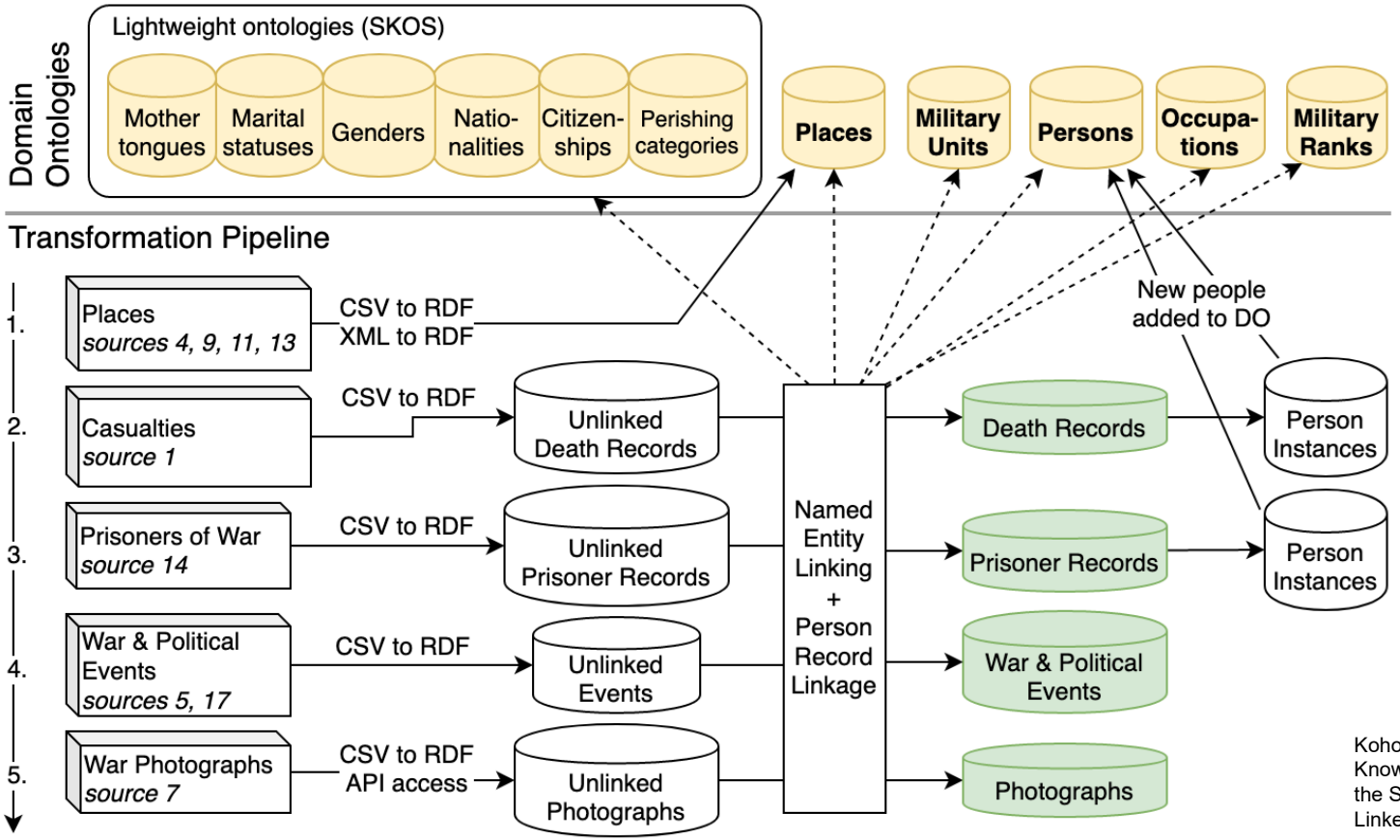
Mikko Koho

WarSampo: Maintenance challenge



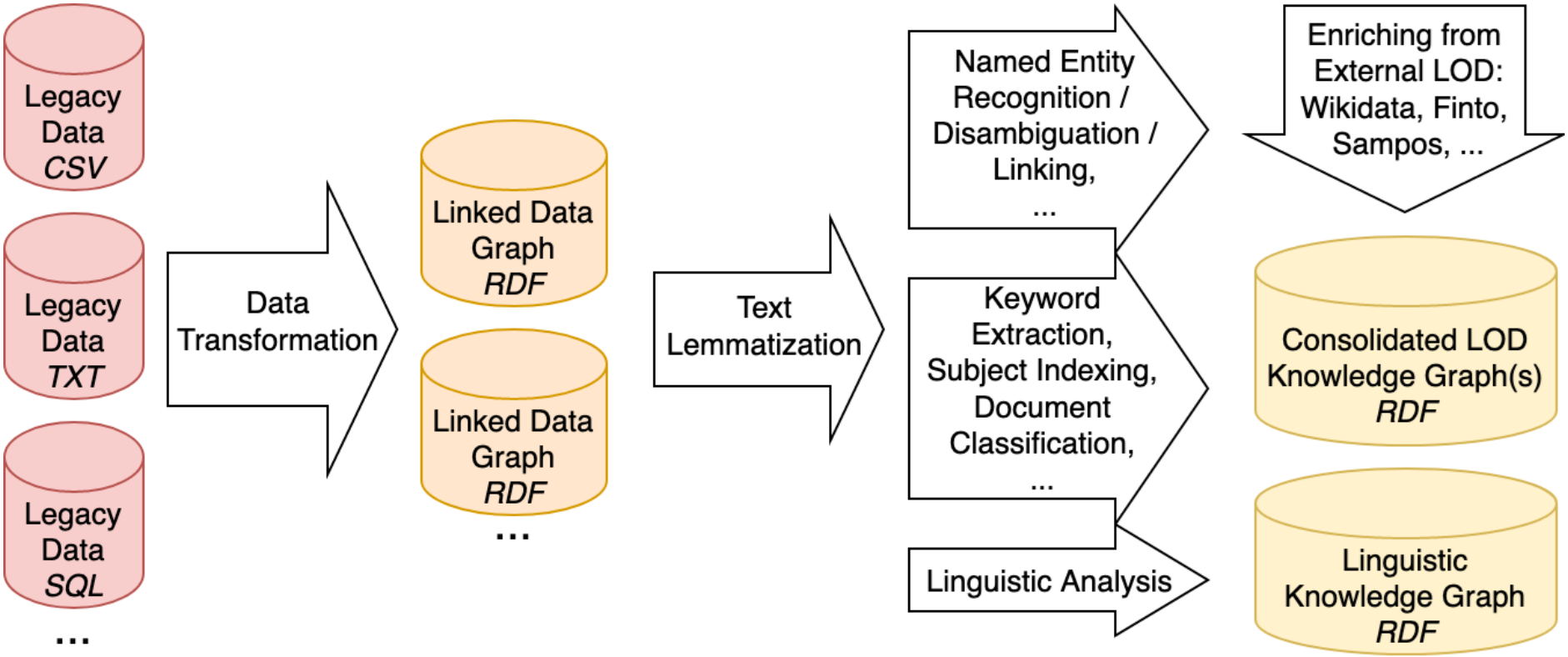
Koho et al.: WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. 2021

WarSampo Data Transformation Pipeline



Koho et al.: WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. 2021

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!