A Linked Open Data Service and Semantic Portal to Study the Assembly Minutes and Prosopography of the League of Nations (1920–1946)

Petri Leskinen[1,2], Eero Hyvönen[2,1], Alexandre Lionnet[3,1], Blandine Blukacz-Louisfert[4], Pierre Etienne Bourneuf[1], Davide Rodogno[1], Grégoire Mallard[1], and Florian Cafiero[3,1]

[1] Geneva Graduate Institute, Center for Digital Humanities and Multilateralism, Geneva, Switzerland
[2] Aalto University, Department of Computer Science, Semantic Computing Research Group (SeCo), Finland
[3] École nationale des chartes, Paris, France
[4] UN Library & Archives Geneva, Geneva, Switzerland

Abstract. This paper presents a new Linked Open Data (LOD) service and a semantic portal on top of it available on the Semantic Web: League of Nations Sampo. This paper shows, how this system can be used for Digital Humanities (DH) research, application development, and can form a basis for a larger LOD intrastructure. In our case, the system is targeted on studying the prosopography and activities of the League of Nations (LoN) (1920–1946), the fore-runner of the United Nations, and is the first concrete step of a larger initiative "Minutes of Multilateralism" for publishing and using a cloud of Knowledge Graphs (KG) about mutually interlinked international organizations in Geneva and beyond. League of Nations Sampo is based on 27 000 pages of minutes of LoN assembly meetings, a prosopographical knowledge graph about some 3100 people mentioned in the minutes, and contextualizing data about the real world. For the first time, this wealth of historical documentation is now openly available as FAIR LOD for DH research and practical application development, as demonstrated by the new League of Nations Sampo portal.

Keywords: knowledge graphs · digital humanities · information retrieval · data analysis

1 Minutes of Multilateralism: studying cross-national organizations

International organizations have produced vast corpora of assembly minutes, plenary debates, committee reports, and voting records. These "minutes of multilateralism" are a core documentary infrastructure for accountability and global governance [3]. Yet scholarship on transparency and record keeping shows that access is uneven and highly shaped by how documents are recorded, registered, and exposed. [21, 9] In practice, many minutes remain locked in scanned volumes or PDFs with heterogeneous metadata and without interoperable, machine-readable representations for searching, browsing, analyzing, and visualizing the contents for Digital Humanities research [5].

Our broader vision is to turn such dispersed materials into a sustainable Linked Open Data (LOD) ecosystem that supports transparency, critical scholarship, and public engagement. In line with the FAIR principles [32] and ongoing work on datafying diplomatic and parliamentary activities [2, 14], this ecosystem should allow users to trace who spoke, on what issues, in which capacities, and how agendas evolved over time, while remaining explicit about uncertainties, biases, and curatorial choices.

The ultimate goal in this work is to create and align with each other minutes data from several related international organizations, with a focus on Geneva-based ones, such as the League of Nations, United Nations, Inter-Parliamentary Union (IPU), and Red Cross. Integrating data from several cross-national organizations, enriched by contextual historical information from related external data sources, such as the Lonsea database[5], Dodis[6], Metagrid[7], and Wikidata, would finally allow for cross-institutional studies based on, e.g., shared prosopographical data about international diplomats and ontologies about politics. For example, the same people have typically been involved in the activities of several organizations[8] and historical events.

In this paper, we present the first concrete step toward this "Minutes of Multilateralism" agenda through the case of the League of Nations (LoN). Its Assembly minutes (1920–1946) offer a rich, bounded laboratory of early multilateralism that makes a valuable source of knowledge for research also by itself. Thanks to the Total Digital Access to the League of Nations Archives (LONTAD) project, [30, 31] these materials are comprehensively digitized and accessible online for humans to read, but not available as Findable, Accessible, Interoperable and Re-usable FAIR data for research and application development: the minutes are available mostly as PDFs with limited structured metadata, which constrains large-scale querying, linking, and computational analysis.

Building on LONTAD, we transformed 27 000 pages of Assembly minutes and related to over 3000 mentioned representatives and other people into a Linked Open Data service and semantic portal League of Nations Sampo. The data was the enriched from external data sources using linked data principles of the Web of Data [8]. This work re-uses and adapts the "Sampo" model of LOD-based semantic portals[9] [11] successfully, used in cultural heritage and

[5] Lonsea: `https://universe.unibas.ch/projects-collaborations/47560`

[6] Dodis: `https://www.dodis.ch/`

[7] Metagrid: `https://www.metagrid.ch/`

[8] According to the Lonsea project data the Greek representative Nicolas Politis (1872–1942) has had nine different roles in eleven organizations

[9] Sampo portal series: `https://seco.cs.aalto.fi/applications/sampo/`

parliamentary contexts, notably the system *ParliamentSampo – Parliament of Finland on the Semantic Web* to publish parliamentary speeches and political prosopography. [15, 28, 20]. The ParliamentSampo experience shows how such infrastructures can enable prosopographical research, longitudinal analyses of political discourse, and civic transparency, and suggests that similar approaches can render multilateral institutions more legible to both experts and the wider public.

The new LoN LOD created, that includes prosopographical and LoN assembly minutes-based data, has been published as linked open data (by CC BY 4.0) with documentation, using W3C best practices[10] (e.g., content negotiation), and as a SPARQL endpoint on the LDF.fi platform[11][16] Also a data dump on Zenodo with a DOI for reference is available[12]. The portal on top of the data service has been opened for feedback from the research community[13] and the software is available under the MIT License in Github[14] for further development.

In the following, Section 2 explicates the research problems and hypotheses of our work, after which the data transformation pipeline for the LOD service is presented (Section 3). The portal to test and demonstrate how the LOD service can be used in practice is presented in Section 5. In conclusion, contributions of the paper are summarized and directions for further research outlined.

2 Research problems and hypotheses

The assembly minutes minutes of international organizations have been openly available as human readable images with minimal metadata but not as FAIR data for searching, browsing, and data-analysis in Digital Humanities research. This paper therefore seeks answers to the following research questions:

1. How to represent assembly minutes documents in a semantic form suitable for Digital Humanities research?
2. How to transform the PDF documents into that format?
3. How to enrich the data with biographical data about the actors involved to facilitate prosopographical research about the international representatives and politicians involved?
4. How to search, browse, and analyze the data as needed by researchers in humanities.

Our research hypothesis is to create an ontological data model of minutes, of the underlying actors, and contextualize these data by knowledge about the real

[10] Best practices for publishing linked data by W3C: `https://www.w3.org/TR/ld-bp/`

[11] League of Nations KG LOD service:: `https://ldf.fi/dataset/lon/`

[12] LoN LOD as data dump : `https://doi.org/10.5281/zenodo.17791448`

[13] League of Nations Sampo portal online (user/password: league/minutes2025): `https://minutes.ldf.fi`

[14] League of Nations Sampo portal software: `https://github.com/SemanticComputing/lon-minutes-web-app`

world. The Linked Data paradigm [8] could be used for representing, interlinking, enriching, and publishing the data. As for the web service model and software tools, the Sampo model and Sampo-UI framework [12] could be adapted and re-used as exemplified by the ParliamentSampo system for publishing minutes of parliamentary discussions [14].

3 Data transformation from primary sources to a LOD service

This section explains the data transformation and LOD publication pipeline created for League of Nations Sampo.

3.1 Data transformation pipeline

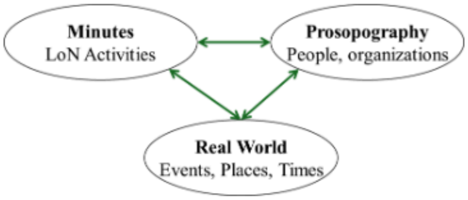


Fig. 1. Interlinked components of the League of Nations Sampo knowledge graph.

The data underlying League of Nations Sampo falls in three parts (cf. Fig. 1). Firstly there is the primary data about the LoN minutes provided by LONTAD as PDF documents with very little metadata. To FAIRify these documents, this material was OCR'd, its named entities were tagged (Subsection 3.1), the entities were disambiguated and linked, and the metadata schema used was populated (Subsection 3.3). Secondly, proposographical data was extracted from seven data sources (listed in Table 1) to facilitate semantic search, browsing, and data analyzes of the Minutes. Here dataset specific Python scripting were used for the RDF transformations. Thirdly, in the same vein, real world knowledge about the places, keywords, and events pertaining the underlying real world was created to give historical context for the Minutes. An event-centric ontological model was adopted for data harmonization, where actors in different roles participate in events that occur in different locations at different times. Finally, the data was published in a SPARQL endpoint that is used directly for DH research and application development, such as the League of Nations Sampo portal.

Table 1. Presence of the 3200 person entities in the interlinked data sources of League of Nations Sampo used for enriching the primary minutes dataset LONTAD.

#	Data Service	People	Domain
1	Lonsea[a]	2653	Database about the history of the League of Nations
2	Wikidata	1682	Knowledge graph and data service underlying Wikipedias
3	VIAF.org	1440	Virtual International Authority File system combining other authority services of national libraries etc.
4	Wikipedia	1296	English or French Wikipedia
5	GND[b]	1181	Integrated authority file system of the the German National Library
6	Metagrid.ch[c]	385	Linking service designed to connect metadata and resources of research projects throughout the humanities and social sciences
7	DODIS[d]	369	Diplomatische Dokumente der Schweiz

[a] Lonsea: `http://www.lonsea.de/`
[b] GND: `https://www.dnb.de/EN/`
[c] Metagrid API service: `https://metagrid.ch/docs/enhancer/`
[d] Dodis: `https://www.dodis.ch/`

3.2 Data transformation: OCR and NER

Due to the complex layouts and sometimes multilingual documents we had to handle, our approach has been twofold (Fig. 2).

First, a subset of pages was labeled for layout regions (columns, main text, marginalia, page headers/footers). These annotations serve to fine-tune a YOLO-based detector [26] that identifies text blocks and, crucially, distinguishes one- and two-column areas. On each page, the detected regions are ordered top-to-bottom and left-to-right within each column, allowing us to reconstruct the correct reading sequence across complex layouts.

In parallel, another subset was transcribed to create OCR ground truth used to benchmark candidate engines; PERO-OCR [23] was selected and applied to the full corpus. On the complete dataset, YOLO layout regions and PERO-OCR line outputs are overlapped to associate each line with its column and structural context. The result is serialized into XML that preserves page structure, segment boundaries, and links to images.

Named Entity Recognition (NER) is then applied using transformer-based multilingual models in the XLM-R family, [4] implemented via Flair [1] and matched against curated authority lists.

3.3 Creating the knowledge graph and SPARQL endpoint

The model schema for the RDF dataset is depicted in Fig.3. The schema is based on CIDOC–CRM standard[15] as well as on the Bio CRM [29] data schema.

[15] CIDOC-CRM model: `https://cidoc-crm.org`

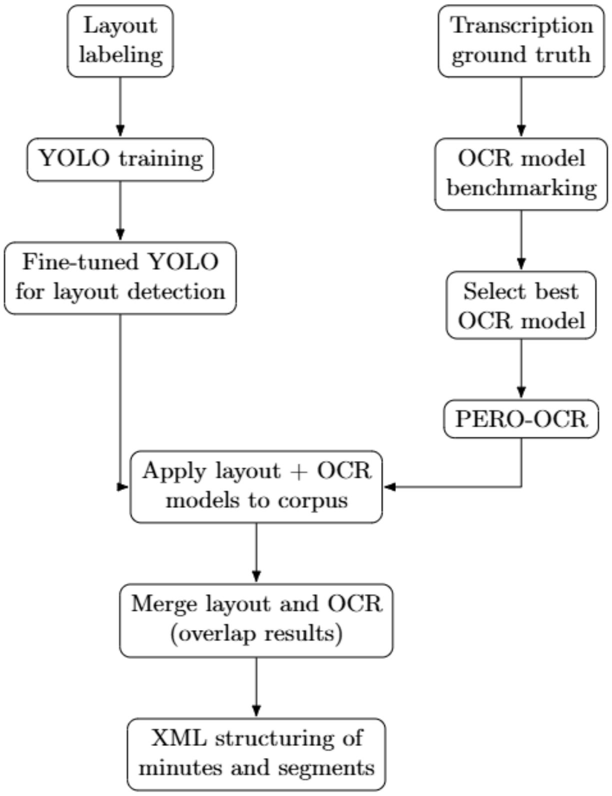

Layout labeling
YOLO training
Fine-tuned YOLO for layout detection
Transcription ground truth
OCR model benchmarking
Select best OCR model
PERO-OCR
Apply layout + OCR models to corpus
Merge layout and OCR (overlap results)
XML structuring of minutes and segments

Fig. 2. End-to-end workflow for layout detection, OCR, and XML structuring of the LON Assembly minutes.

A similar schema has previously been used, for example, for parliamentary data [20]. The model facilitates to modeling and enriching actor data with related events, such as holding a specific position or belonging to a specific organization. The resources of the minutes (`:Minute`) and the references (`:Reference`) produced in the NER process form the core of the data. Each (`:Minute`) contains the fields of the speech text both as a plain text (`:content`) and in HTML format containing the tagging for showing it in an online browser (`:html`). In the RDF different types of named entities are modeled using classes: 1) (`crm:E21_Person`) for individual people, 2) (`crm:E74_Group`) for organizations, 3) (`crm:E53_Place`) for locations, 4) (`crm:E52_Time-Span`) for temporal expressions, and 5) (`skos:Concept`) for miscellaneous references. Furthermore, the schema of Bio CRM is applied to model temporal group memberships (`crm:E7_Activity`) with different roles (`biocrm:Actor_Role`).

Besides the minute speeches, the recognized individual people are the second main focus in the data publication. In the NER results for people there were basically three types of references:

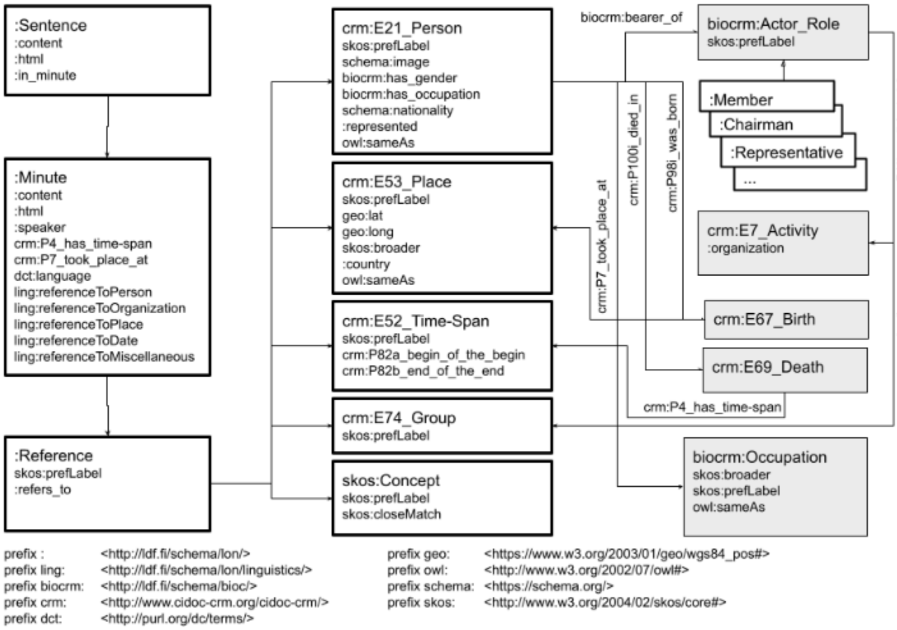


Fig. 3. RDF data schema

1. Mamed individuals: "M. Andreas Oldenburg" or "général Nemours
2. Individuals mentioned by title: "Danish representative" or "Secrétaire général"
3. Anonymous groups or individuals: "technical officers", "aggressor" or "women".

In Named Entity Linking (NEL) the main focus has been in disambiguating and linking the entities of the first type to external databases. In average an individual was referred in the minutes texts using 2.59 different literal forms. As an example the Salvadoran representative José Gustavo Guerrero (1876–1958) was referred in 24 different ways (Mr. Guerrero, M. Gustavo Guerrero, Dr. J. Guerrero, J. G. Guerrero, etc.) including the ones containing typos.

The references containing person names or geographical locations are disambiguated against data extractions from external databases, e.g., League of Nations Search Engine (Lonsea)[16] and Wikidata. Furthermore, Wikidata identifiers and Metagrid API service[17] are used to add linkage to databases such as GND, VIAF, and DODIS. Finally, the prosopographical person data is enriched with name variations, places and times of birth and death, nationalities, occupations, and images. The ontology of places represents the geographical locations mentioned in the minutes, the places related to the biographical details of the people, and the nationalities. The hierarchy in the ontology is wide-spread,

[16] League of Nations Search Engine: `http://www.lonsea.de/` NB. the site might be occasionally unavailable

[17] Metagrid widget api reference: `https://metagrid.ch/docs/widget/reference/`

reaching from the level of a single building to towns, countries, and finally continents.

In the conversion process the named entity linking was performed using Python modules SPLink[18], Nameparser[19], Deep-translator[20], and PolyFuzz[21] taking into consideration the time of speech as well as the other named entities mentioned in the same minutes or in their sentences, e.g. locations and nationalities. The most of the false or missing links were caused by

1. not having the biographical data available in our chosen external sources,
2. name variations or differing practices in spelling (specially with East Asian person names), or
3. multiple candidates, e.g., in cases when only the family name of a person is mentioned.

3.4 Enriching the data from external sources

After transforming the LONTAD minutes data into RDF form it was enriched from related data sources by data linking, This was done by aligning person, organization, and place entities with corresponding data entries in seven external data services listed in Table 1. The table quantifies, as an example, that most of the 3200 key people in LONTAD are also present in one or more related web services. This means that lots of additional data about the people could be extracted into the League of Nations Sampo KG, making the data richer than in LONTAD.

As the data comes from several sources, provenience information about the aggregated triples are included in the KG. This metadata is useful for making the data transparent to the end user. In our case, 1) provenience data can be for verifying results in data analyses from the primary sources, 2) for showing data sources to the end user in application user interfaces, and 3) also to find automatically possible inconsistent pieces of shared data across the contributing data services, as suggested and shown in [19].

3.5 LOD service

The enriched KG was published on the Linked Data Finland platform LDF.fi[22] [17] using the best publishing practices of the W3C, including a SPARQL endpoint and related services, such as content negotiation, resolving URIs, RDF browsing etc. [17]. LDF.fi also allows publication of schemas alongside the actual data as well as automatic data documentation[23]. In LDF.fi each dataset is described

[18] SPLink: `https://moj-analytical-services.github.io/splink/index.html`
[19] Nameparser: `https://pypi.org/project/nameparser/`
[20] Deep-translator: `https://pypi.org/project/deep-translator/`
[21] PolyFuzz: `https://maartengr.github.io/PolyFuzz/`
[22] Linked Data Finland platform: `https://ldf.fi`
[23] Using pyLODE tool `https://github.com/RDFLib/pyLODE` based on LODE

using VoID[24], where a rating of 1–8 stars can be given, extending Tim Berners-Lee's 5-start model [16] (the addtional stars are for publishing the schemas, validating the data against the schemas, and for truthfulness of the data). Based on the VoID metadata, a homepage for the dataset with a SPARQL endpoint, associated LOD services, and instructions for re-using the data are automatically created. LDF.fi is part of a national Finnish LOD infrastructure [13].

4 Using the LOD service for Digital Humanities research

The LOD service is useful for several purposes. 1) It can be used directly by scripting for DH research. 2) It provides a basis on which applications such as portals can be built. 3) It can be used to enrich other related KGs as envisioned by the Minutes of Multilateralism agenda. In this section, demonstrational illustrations of the use case (1) are given, (2) is discussed in Section 5, and (3) in conclusions as a topic for further research.

A common way to use an LOD endpoint is by a Google Colab or Jupyter notebook or in R-Studio. One can first query the available endpoint for the needed information, and then visualize the results using datasheets, time series, charts, or networks. An example in our case study is given in this notebook[25]. This notebook was created using Python modules, such as SPARQLWrapper for database query, and Pandas, NumPy, and Matplotlib for data analysis and visualization.

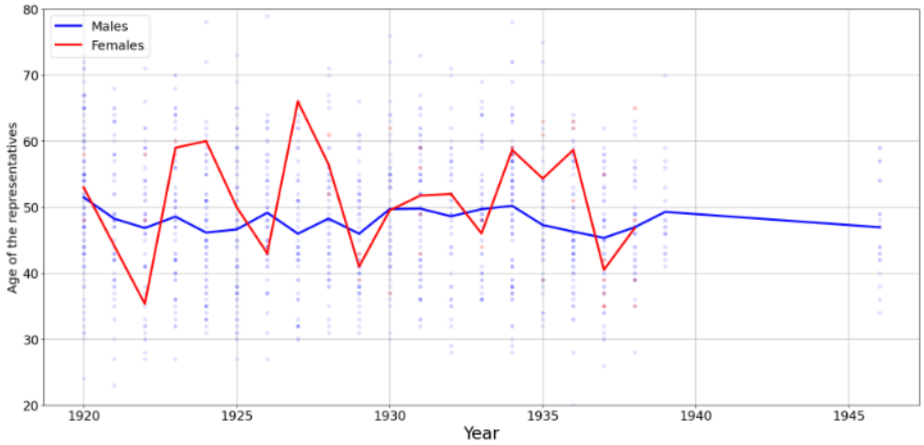


Fig. 4. Ages of the General Assembly Representatives

[24] VoID Vocabulary: `https://www.w3.org/TR/void/`

[25] Google Colab example: https://colab.research.google.com/drive/1v8fhoO1I5McMGhtHKuNo5fD4i4M8nBRd?usp=sharing

An output created the example notebook is depicted in Figure 4. The continuous lines show the average age of the General Assembly representatives during the years 1920–1946 for males and females with blue and red, respectively. Since the total number of female representatives is very low, 46 out of the total 1115, even a minor change in the number of members affects the average. The dots in the background illustrate the ages of individual representatives.

5 League of Nations on the Semantic Web: the portal

This section presents first the generic portal UI model supported by the Sampo-UI framework and then illustrates by examples how the portal is used. The examples demonstrate some of the benefits of the Sampo-UI model and data linking in contrast to traditional legacy systems, typically based on a single data-silo and without FAIR data available for searching, browsing, and analyzing the contents.

5.1 Portal model for minutes data

To test and demonstrate the approach above, a web application League of Nations Sampo was built by using the Sampo-UI framework [18, 25][26] and the domain specific ParliamentSampo framework [15] on top of it. The idea of the "Sampo framework" is to take an existing Sampo in a domain of interest (here speeches and documents related to assembly meetings), with its ready to use user interface (UI) model and knowledge graph as a starting point. The UI specification is a set of JSON specifications that are then modified *declaratively* for the new UI. In addition, the underlying SPARQL queries are adapted to access the more or less different KG of the new Sampo. This approach allows for extremely rapid prototyping and software development, if the framework used (in our case ParliamentSampo) is mostly fit for the new purpose. Re-using an existing framework requires only modest programming skills, but for a more experienced developer, also the framework can be modified and extended with new components and functionalities as needed.

The Sampo-UI consists of two main components: (1) a client-side interface built using the well-established React[27] and Redux[28] libraries and (2) a Node.js[29] back-end developed with the Express[30] framework. This framework allows developers to reuse components, such as faceted search, data tables, data analyses, and visualisations through a specification-based configuration.

An ambitious idea behind Sampo-UI is to provide a generic approach for accessing KGs of different kinds in *any* SPARQL endpoint on the Web, based

[26] Sampo-UI home: https://seco.cs.aalto.fi/tools/sampo-ui/; Github: https://github.com/SemanticComputing/sampo-ui

[27] React webpage: https://reactjs.org

[28] Redux webpage: https://redux.js.org

[29] Node.js webpage: https://nodejs.org/en

[30] Express webpage: https://expressjs.com

on the core concepts of the Resource Description Framework (RDF)[31], the foundation of the Semantic Web: Classes, Properties, and Individuals (Instances). (Cf., e.g., work on the Nobel Price Sampo [6] at the University of Latvia, based on an external SPARQL endpoint provided by the Nobel Foundation in Sweden.) Each class, i.e., entity type of interest, can have an *application perspective*, which is a faceted semantic search view for filtering and exploring individuals, i.e.,instances, of the perspective class. Each perspective involves two main types of UI components:

1. One for faceted search [7], earlier called view-based search [24], to filter instances of the perspective class using semantic properties and associations of the class, including dynamically updated hit counts to direct search and prevent dead-end 'no hits' situations in browsing. Sampo-UI integrates this search paradigm into the world of the Semantic Web ontologies, as suggested in [10].
2. Visual components on separate tabs for the filtered search results through tools such as maps, timelines, networks, and tables for semantic exploration.

5.2 League of Nations Sampo portal

The LoN data publication can be used for two main purposes.

1. For straight-forward DH research and analyses on top of the SPARQL endpoint, using tools such as the YASGUI editor with visualizations [27] and Google Colab for scripting.
2. For developing applications, such as portals, that can be used without programming skills.

As a use case of the latter option, this section presents the portal League of Nations Sampo in order to demonstrate and evaluate the feasibility of the vision and agenda of "Minutes of Multilateralism" presented in Section 1.

The landing page of the portal (cf. Fig. 5) presents three application perspectives for searching Minutes, People, and Places. Below them shortcut links to example searches and visualizations are presented: one to the minutes where the ownership question of the Åland Islands between Finland and Sweden was discussed (in the Minutes perspective) and one to a visualization about people mentioned in the minutes who died in Geneva (in People perspective), depicted in Fig. 6. On the upper bar, links to perspectives, instructions, project information, and to a feed-back channel are always readily available.

In Fig. 6, 11 facets are shown on the left with the facets Nationality and External databases opened. The result set includes 23 people listed on the right with images from Wikidata and links to metadata values for more information, especially to the *instance pages* (homepages) of each individual. In Sampo-UI each individual instance can be associated with a "homepage" where information about the individual is gathered by data linking and reasoning. In this case, the

[31] RDF of W3C: `https://www.w3.org/RDF/`

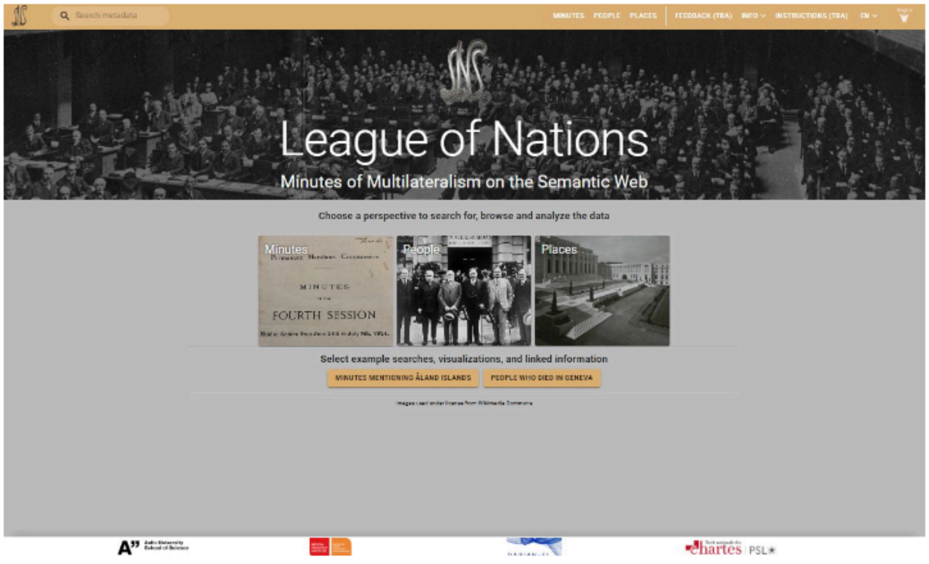
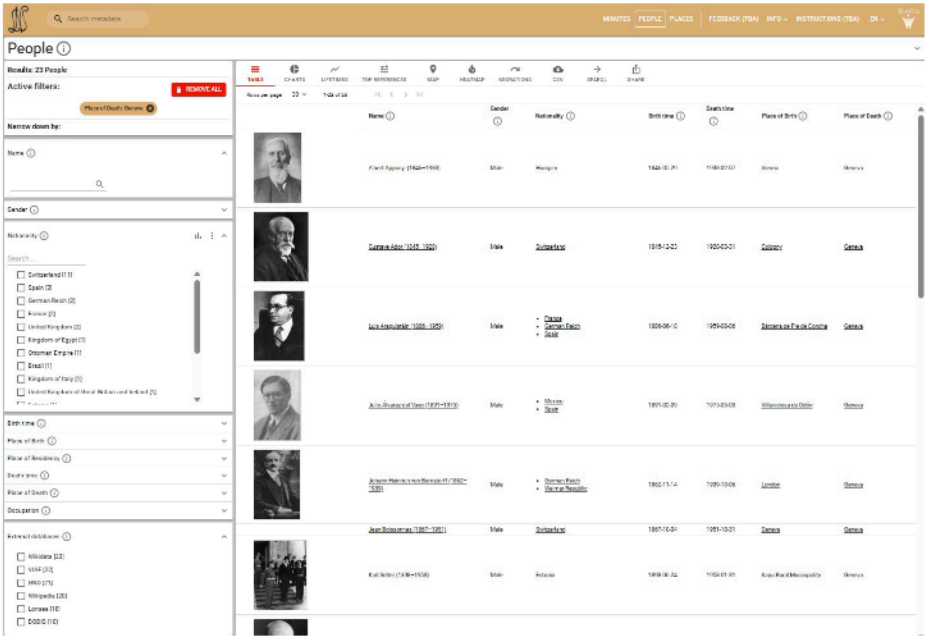

Search metadata
MINUTES
PEOPLE
PLACES
FEEDBACK (TBA)
INFO
INSTRUCTIONS (TBA)
EN
League of Nations
Minutes of Multilateralism on the Semantic Web
Choose a perspective to search for, browse and analyze the data
Minutes
People
Places
MINUTES
FOURTH SESSION
Select example searches, visualizations, and linked information
MINUTES MENTIONING ÅLAND ISLANDS
PEOPLE WHO DIED IN GENEVA
Images used under license from Wikimedia Commons
chartes
PSL

Fig. 5. Landing page of League of Nations Sampo



Name	Gender	Nationality	Birth time	Death time	Place of Birth	Place of Death
Albert Apponyi (1846–1933)	Male	Hungary	1846-05-29	1933-02-07	Vienna	Geneva
Gustave Ador (1845–1928)	Male	Switzerland	1845-12-23	1928-03-31	Cologny	Geneva
Luis Araquistáin (1886–1959)	Male	• France • German Reich • Spain	1886-06-10	1959-08-08	Bárcena de Pie de Concha	Geneva
Julio Álvarez del Vayo (1891–1975)	Male	• Mexico • Spain	1891-02-09	1975-05-03	Villaviciosa de Odón	Geneva
Johann Heinrich von Bernstorff (1862–1939)	Male	• German Reich • Weimar Republic	1862-11-14	1939-10-06	London	Geneva
Jean Boissonnas (1867–1951)	Male	Switzerland	1867-10-04	1951-10-31	Geneva	Geneva
Karl Selter (1898–1958)	Male	Estonia	1898-06-24	1958-01-31	Kopu Rural Municipality	Geneva

Fig. 6. Application perspective for people mentioned in the LoN minutes

the individuals represent a wide cross-national spectrum of nationalities and the data has been harvested from a variety of data sources. These sources are shown and available for search on the open External databases facet on the left low corner, and include Wikidata, Virtual International Authority File service VIAF.org provided by OCLC in the United States, Gemeinsame Normdatei GND register of the German National Library, Wikipedia, Lonsea, and Dodis. For example, the excessive Lonsea database knows 18 out of the 23 people, which demonstrates the added value earned by the linked data approach for enriching data.

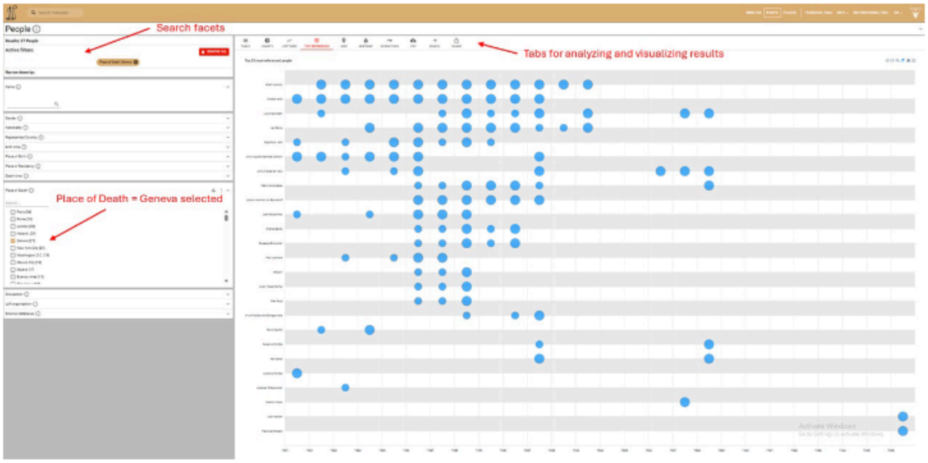

Search facets
Tabs for analyzing and visualizing results
Place of Death = Geneva selected

Fig. 7. Visualization tab for illustrating the amount of mentions of people in the minutes on a timeline. Tab TOP REFERENCES is selected.

Above the search result, a list of data-analytic visualization tabs are available for Digital Humanities research in addition to the default TABLE view: CHARTS, LIFETIMES, MAP, HEATMAP, and MIGRATIONS, In addition the CVS tab can be used to import the results into a CSV table for external tools (e.g., R) for further analysis. The SPARQL tab opens a YASGUI interface for querying the SPARQL endpoint and SHARE tab to share the URLs of the visualization. For example, Fig. 7 depicts on the timeline of the TOP REFERENCES tab how often the people who died in Geneva are mentioned in the minutes.

In Fig. 8 shows on the MIGRATIONS tab a visualization of how the international LoN people have moved from their place of birth (blue end of an arc) to place of death (red end of the life line). Clicking on an arc open an popup window showing the related locations and the involved people. In the example case two people were born in Kolkata and died in Kensington, UK. Because it was still colonial era, very few of the representatives originate from Africa.

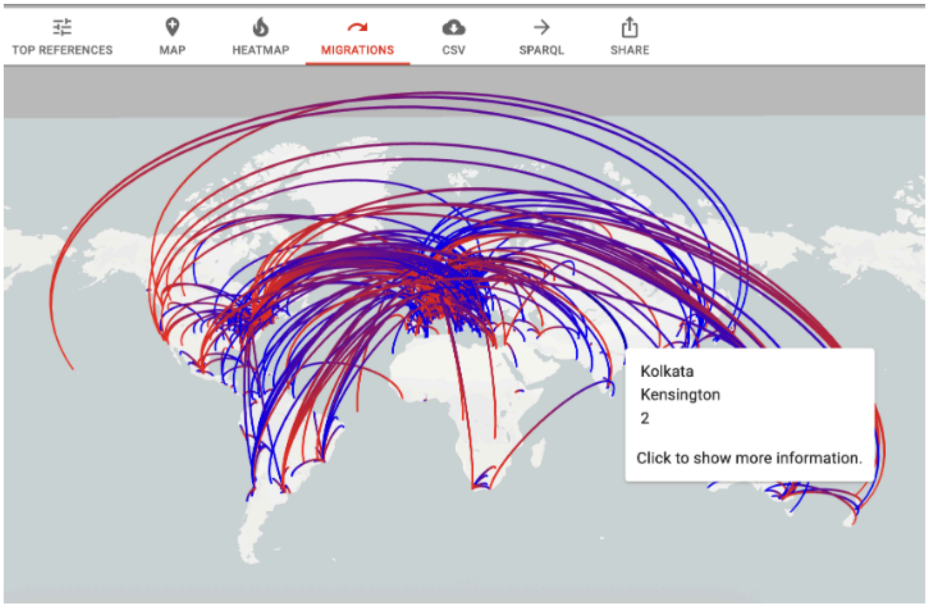


Fig. 8. MIGRATIONS tab for illustrating movements of LoN people on a map.

6 Conclusions

This paper presented how an existing Sampo framework with its portal UI design on parliamentary speeches and prosopography could be re-used for a related system on assembly minutes of an international organization, the League of Nations. Based on the OCR'd minutes, NER data, and related external data sources for enriching the data, the first version of the League of Nations Sampo data service and portal could be created in a couple of months, because the ParliamentSampo framework and the Sampo infrastructure could be adapted and re-used in this new analogous application case.

The new resource is available openly on the Web as a LOD service, as data dumps, and as a semantic portal for DH research, application development, and as a basis for an infrastructure about international organizations, diplomats, and their psosopography. The system can be used for enriching further related datasets, such as the General Assembly Minutes of the United Nations (UN), Inter-Parliamentary Union IPU, and Red Cross. By moving on from publishing PDF documents, as customary in current legacy systems, to publishing also the underlying data as FAIR KGs in RDF form adds value substantially to these historically important documents.

As a proof of concept, this paper demonstrated by examples how the data service can be used directly for DH research and for application development. Work is currently underway for extending the infrastructure next to the General

Assembly Minutes of the United Nations, the follow-up organization of LoN after the Second World War.

A key practical challenge in systems such as League of Nations Sampo is sustainability: how to maintain the data service and the portal in the future. For example, how are possible changes propagated to League of Nations Sampo if data in the primary sources are updated or what happens if the services are terminated? In order to minimize dependency risks pertaining to the primary data sources, League of Nations Sampo copies essential data into its own KG. If the primary sources provide a functioning API, such as the SPARQL endpoint in the case of Wikidata, it can be used used for dynamic updates and for providing back links for further information in the original data owner's website, such as a Wikipedia and Dodis.ch. In our case, many of the data sources are not expected to change much, and the case is in this respect easier than, e.g., ParliamentSampo where new speeches may come in on a daily basis and the prosopography changes substantially after every new election of the Parliament. We anticipate that League of Nations Sampo has the potential of becoming a new primary source of LoN data in the larger Minutes of Multilateralism infrastructure to be maintained directly in native RDF form. For this purpose, some earlier Sampos in use, such as BookSampo and OperaSampo, have used the SAHA editor on top of the SPARQL endpoint [22].

To addresss the sustainability challenge to start with, the data service and portal software of League of Nations Sampo are included as a demonstrator in the FIN-CLARIAH/DARIAH-FI infrastructure programme (2022–2029). Furthermore, the data service and portal are available as Docker containers[32], which makes it very easy to re-install the services if needed on different server platforms or locally on a personal computer, including compatible versions of the underlying software packages.

Acknowledgments. We thank the LONTAD archives project, UN Library & Archives Geneva, Lonsea project, Dodis, and Metagrid for fruitful collaborations. Support of the Finnish Digital Humanities research infrastructure initiative FIN-CLARIAH/DARIAH-FI funded by the Research Council of Finland and NextConnectionEU is acknowledged. Computational resources of CSC – IT Center for Science have been used in the research.

[32] Docker containers: `https://www.docker.com/`

Bibliography

[1] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59 (2019)

[2] Cafiero, F.: Datafying diplomacy: How to enable the computational analysis and support of international negotiations. Journal of Computational Science **71**, 102056 (2023). `https://doi.org/10.1016/j.jocs.2023.102056`

[3] Cafiero, F., Cointet, J.P., Mallard, G.: Digital accountability can relegitimate multilateralism. HAL preprint (2025), `https://hal.science/hal-05396546v1`, preprint, HAL Id: hal-05396546v1. Accessed December 3, 2025

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451 (2020)

[5] Drucker, J.: The Digital Humanities Coursebook. An Introduction to Digital Methods for Research and Scholarship. Routledge (2021)

[6] Grislis, N.K., Čerāns, K., Grasmanis, M., Rantala, H., Hyvönen, E.: How to add a user interface on top of an external sparql endpoint: Case nobel prize sampo. In: SEMANTiCS-PDWT 2025, Posters, Demos, Workshops, and Tutorials at SEMANTiCS 2025. vol. 4064. CEUR Workshop Proceeding (2025), `https://ceur-ws.org/Vol-4064/PD-paper17.pdf`

[7] Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., Lee, K.P.: Finding the flow in web site search. CACM **45**(9), 42–49 (2002)

[8] Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool, Palo Alto, California (2011), `http://linkeddatabook.com/editions/1.0/`

[9] Heikkonen, S.: Transparency materialised: how registers can regulate access to documents. European Law Open **3**(1), 1–24 (2024). `https://doi.org/10.1017/elo.2024.7`

[10] Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology techniques to view-based semantic search and browsing. In: The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004). Springer–Verlag (2004)

[11] Hyvönen, E.: Digital humanities on the semantic web: Sampo model and portal series. Semantic Web **14**(4), 729–744 (2023). `https://doi.org/10.3233/SW-223034`

[12] Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series. Semantic Web **14**(4), 729–744 (2022). `https://doi.org/10.3233/SW-223034`

[13] Hyvönen, E.: How to create a national cross-domain ontology and linked data infrastructure and use it on the semantic web. Semantic Web (2024), https://doi.org/10.3233/SW-243468, dOI: 10.3233/SW-243468

[14] Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Leal, R., La Mela, M., Tuominen, J., Poikkimäki, H., Rantala, H.: Publishing and using parliamentary linked data on the semantic web: Parliamentsampo system for parliament of finland. Semantic Web **16**(1) (2025). https://doi.org/10.3233/SW-243683

[15] Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Leal, R., La Mela, M., Tuominen, J., Poikkimäki, H., Rantala, H.: Publishing and using parliamentary linked data on the semantic web: Parliamentsampo system for parliament of finland. Semantic Web **16**(1) (2025), dOI: 10.3233/SW-243683

[16] Hyvönen, E., Tuominen, J.: 8-star linked open data model: Extending the 5-star model for better reuse, quality, and trust of data. In: Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems (SEMANTiCS 2024). vol. 3759. CEUR Workshop Proceedings (9 2024), https://ceur-ws.org/Vol-3759/paper4.pdf

[17] Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events. pp. 226–230. Springer–Verlag (May 2014). https://doi.org/10.1007/978-3-319-11955-7_24

[18] Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web **13**(1), 69–84 (2022). https://doi.org/10.3233/SW-210428

[19] Leskinen, P., Ahola, A., Rantala, H., Tuominen, J., Hyvönen, E.: Consistency checking in a cloud of interlinked cultural heritage knowledge graphs – first results of using the samposampo data service and portal (October 2025), submitted for review

[20] Leskinen, P., Hyvönen, E., Tuominen, J.: Members of parliament in finland knowledge graph and its Linked Open Data service. In: Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems (SEMANTiCS 2021). pp. 255–269. IOS Press (2021). https://doi.org/10.3233/SSW210049

[21] Casadesús de Mingo, A., Cerrillo-i Martínez, A.: Improving records management to promote transparency and prevent corruption. Government Information Quarterly **35**(4), 624–632 (2018). https://doi.org/10.1016/j.giq.2018.09.008

[22] Mäkelä, E., Hyvönen, E.: Sparql saha, a configurable linked data editor and browser as a service. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. pp. 434–438. Springer-Verlag (May 2014), https://doi.org/10.1007/978-3-319-11955-7_62

[23] Novotný, V., Horák, A., et al.: PERO-OCR: An open-source optical character recognition system. https://pero-ocr.fit.vutbr.cz (2022), accessed 2025-11-10

[24] Pollitt, A.S.: The key role of classification and indexing in view-based searching. Tech. rep., University of Huddersfield, UK (1998), http://www.ifla.org/IV/ifla63/63polst.pdf

[25] Rantala, H., Ahola, A., Ikkala, E., Hyvönen, E.: How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework. In: VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. vol. 3508. CEUR Workshop Proceedinbgs (2023), https://ceur-ws.org/Vol-3508/paper3.pdf

[26] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016)

[27] Rietveld, L., Hoekstra, R.: The YASGUI family of SPARQL clients. Semantic Web **8**(3), 373–383 (2017). https://doi.org/10.3233/SW-150197

[28] Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., La Mela, M., Hyvönen, E.: Plenary debates of the parliament of finland as Linked Open Data and in Parla-CLARIN markup. In: Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021). OASIcs, vol. 93, pp. 1–17. Schloss Dagstuhl–Leibniz-Zentrum für Informatik (2021), https://drops.dagstuhl.de/opus/volltexte/2021/14544/

[29] Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: BD2017 Biographical Data in a Digital World 2017, Proceedings. pp. 59–66. CEUR WS Proceedings (2018), http://ceur-ws.org/Vol-2119/paper10.pdf

[30] UN Library & Archives Geneva: LONTAD: Total digital access to the league of nations archives. https://libraryresources.unog.ch/lontad (2022), digitization project (2017–2022) providing comprehensive online access to the League of Nations archives

[31] Wells, C.M.: Total digital access to the league of nations archives: Digitization, digitalization, and analog concerns. In: Archiving Conference. vol. 16, pp. 12–16. Society for Imaging Science and Technology (2019)

[32] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18