# Using Linked Data for Prosopographical Research of Historical Persons: Case U.S. Congress Legislators

Goki Miyakita[1], Petri Leskinen[2], and Eero Hyvönen[2,3]

[1] KMD Research Institute, Keio University, Japan
[2] Semantic Computing Research Group (SeCo), Aalto University, Finland
[3] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://seco.cs.aalto.fi, http://heldig.fi

**Abstract.** This paper shows how biographical registries can be represented as Linked Data, enriched by data linking to related data sources, and used in Digital Humanities. As a use case, a database of 11 987 historical U.S. Congress Legislators in 1789–2018 was transformed into a knowledge graph. The data was published as a Linked Data service, including a SPARQL endpoint, on top of which tools for biographical and prosopographical research are implemented. A faceted browser named *U.S. Congress Prosopographer* with visualization tools for knowledge discovery is presented to provide new insights in political history.

**Keywords:** Linked Data, Visualization, Biography, Prosopography, Digital Humanities, American History

## 1 Prosopographical Method

Person registries and biographies are widely used to document and describe life stories of historical people, with the aim of getting a better understanding of their personality, actions, and motivations in history. In *biography* [17] the focus is on individual protagonists, while in *prosopography* [21] life histories of groups of people are studied in order to find out some kind of commonness or average in them. Here persons are studied in the context of communities. The prosopographical research method [21, p. 47] consists of two steps. First, a target group of people is selected that share desired characteristics for solving the research question at hand. Second, the target group is analyzed and compared with other groups to solve the research question.

This paper shows how the prosopographical method can be used in practice in Digital Humanities by presenting a tool and application based on the Linked Data (LD) paradigm [5] in the Cultural Heritage (CH) domain [7]. It is shown how faceted search and data visualization tools can be integrated with a SPARQL endpoint allowing the end user to 1) filter out target groups of people, and 2) then to study them. A key novelty of this paper is the idea to support comparing analyses and visualizations based on different target subgroups. The paper extends substantially our earlier short four page paper [14] that focuses on the application demonstrator.

As a use case, a database about the United States Congress Legislators[4,5] were used. We pulled and linked two different datasets: 1) Dataset of the members of the United States Congress and 2) Dataset based on ICPSR ID[6] accompanying Congress numbers[7], as a basis. It contains biographical records of 11 987 persons who served in the U.S. Congresses from the 1st (1789) to the 115th (2018) one. We converted and extracted the data into RDF, and developed a SPARQL compliant data service and an online application named *U.S. Congress Prosopographer*[8] to complement both quantitative and qualitative inquiry in American political history.

The paper presents first a data model we used in representing person registries and a data service for publishing the registry data as Linked open Data. After this, the tools for using the data service for biography and prosopography are presented, and a set of use cases are described. In conclusion, contributions of the work are summarized in relation to related works, and directions for further research are suggested.

## 2 Data Model and Linked Data Service

**Data Model** The ontology model representing people and their biographical information is based on the schema.org vocabulary[9] [3]. Legislators are modeled as instances of the class `schema:Person` with properties expressing their biographical data. Places are represented as instances of `schema:Place`. The data model of schema.org is extended by additional properties and classes in the domain specific namespace[10] of the underlying data service (to be described below). To illustrate the model, a example of data resources in the Turtle format[11] is given below:

```
@prefix :       <http://ldf.fi/congress/> .  # Domain name space
@prefix schema: <http://schema.org/> .
@prefix wd_ent: <http://www.wikidata.org/entity/> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix skos:   <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix geo:    <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix dbr:    <http://dbpedia.org/resource/> .

:p10079  a      schema:Person ;   # A person instance
   schema:familyName    "Truman" ;
   schema:givenName     "Harry" ;
   schema:gender        "Male" ;
   schema:birthDate     "1884-05-08"^^xsd:date ;
   schema:birthPlace    wd_ent:Q572172 ;
   schema:deathDate     "1972-12-26"^^xsd:date ;
   schema:deathPlace    wd_ent:Q41819 ;
   :bioguide_id         "T000387" ;
   :dbpedia_id          dbr:Harry_S._Truman ;
   :govtrack_id         "410956" ;
```

---

[4] https://github.com/unitedstates/congress-legislators

[5] http://k7moa.com

[6] The Inter-university Consortium for Political and Social Research (ICPSR) ID number

[7] https://www.senate.gov/reference/Years_to_Congress.htm

[8] https://semanticcomputing.github.io/congress-legislators

[9] http://schema.org/docs/schemas.html

[10] http://ldf.fi/congress/

[11] https://www.w3.org/TR/turtle/

```
    :icpsr_id              "9487" ;
    :type                  "Senator" ;
    :wikidata              wd_ent:Q11613 ;
    :wikipedia_id          "Harry S. Truman" ;
    rdfs:comment           "Harry S. Truman (May 8, 1884 ..."@en ;
    schema:description     "Harry S. Truman (May 8, 1884 ..."@en ;
    schema:hasOccupation   "statesman" , "politician" ;
    schema:image           "HarryTruman.jpg" ;
    schema:memberOf        "Democrat" ;
    schema:state           "MO" ;
    skos:prefLabel         "Truman, Harry" .

wd_ent:Q41819  a           schema:Place ;    # A place instance
    rdfs:label             "Kansas City"@en ;
    schema:containedInPlace  wd_ent:Q127238 ;
    geo:lat                39.05 ;
    geo:long               -94.583333333 .
```

Here the *Harry Truman* resource `:p10079` is represented as an instance of the class `schema:Person`. All basic biographical data (family name, gender, given name, etc.) are modeled using the schema.org namespace, and all the data relating to his career are in the domain namespace. The resources can be linked to external databases or services, such as DBpedia, Wikidata, Wikipedia, and Twitter, for more information. There are 11 987 legislator resources like this in the knowledge graph.

The data example also shows the place resource of *Kansas City*, `wd_ent:Q41819`. The place ontology data in our domain was extracted from DBpedia[12]. The dataset contains 7865 place resources, including all places mentioned in the United States Congress Legislators data. Each place entry contains the place name, the latitude and longitude, and a link to a larger scale place containing the current place, a linkage that constructs a topological hierarchy. Property `:icpsr_id` is used to identify and classify each person according to the Congress(es) (s)he served/serves.

**Data Service** The data is available as a linked open data service at the Linked Data Finland platform[13] [8] in an open SPARQL endpoint[14] with resolvable URIs, using the W3C Linked Data publishing principles and best practices [5]. For example, the URI http://ldf.fi/congress/p10079 refers to Harry Truman (1882–1972), and can be used for retrieving the related RDF data or for Linked Data browsing depending on the need and HTTP protocol header data used.

The data in the service contains altogether ca 830 000 triples, 790 000 in the people graph and 40 000 in the place graph.

## 3   Supporting Biographical and Prosopographical Research

This section describes the functionalities of the tool *U.S. Congress Prosopographer* from the end-user's perspective. To support the first step in the prosopographical method, a faceted search application view is provided for finding legislators of interest and their groups. To support the second analysis step in the prosopographical method, the tool contains separate application views for 1) visualizing the target group on a map, 2) analyzing it using statistics, and 3) for making comparisons between two target groups. In

---

[12]http://dbpedia.org

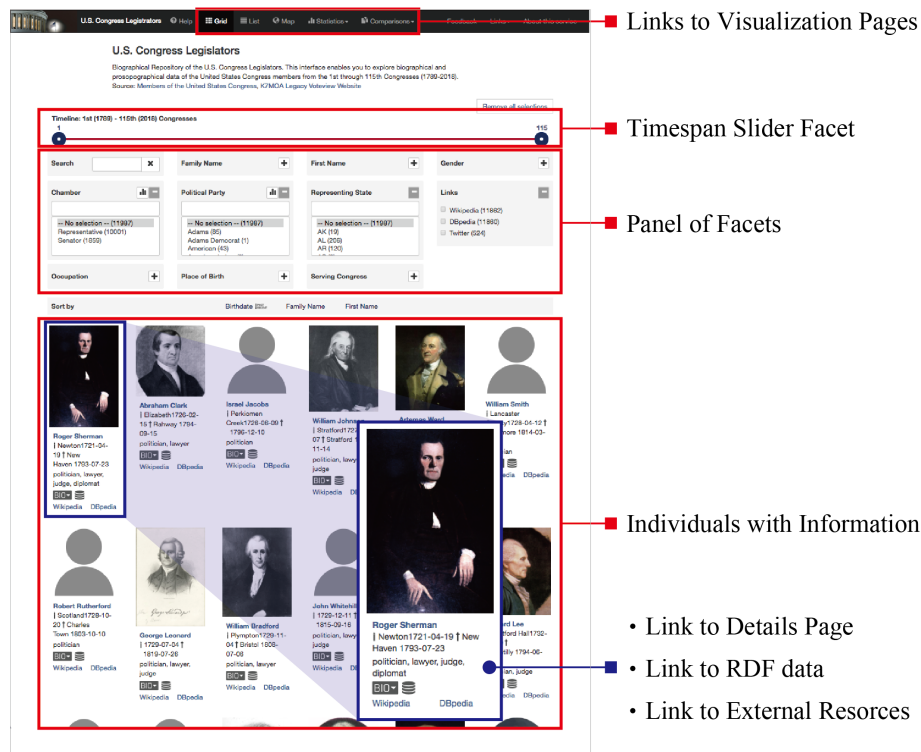[13]http://ldf.fi

[14]http://ldf.fi/congress/sparql

**Fig. 1.** Faceted Search View (Search Interface)

below, these features are explained in more detail. All application views can be selected from the menu bar of the system using corresponding link buttons that open the selected application view.

**Faceted Search View** Fig. 1 depicts the main faceted search interface of the tool. Its upper part allows the user to filter the target group of interest in the prosopographical method, based on filters and hierarchical facet ontologies shown on the top. The lower part visualizes the target group retrieved as a matrix grid. Its cells contain an image of the person (if available) and below it metadata about the person with links to additional data sources. The matrix can alternatively be viewed in list form by clicking the link "List" on the menu bar instead of "Grid". The use case of this application view is to find groups of persons of interest, and then by selecting one of them, to get more information about him/her on a Details Page (a kind of "home page") (cf. Fig. 1). For biographical research, the Details Page of each person entry contains a description of the person, based on the original database, with additional data links. The description is also enriched by a short textual biograph extracted from Wikidata, when such an article is available.

In the data filtering part, there is first a timespan slider by which the range of congresses (1st–115th) in focus can be constrained. Below that there is generic search box filter, like in Google basic search box. After this comes facets for faceted search.

The idea in faceted search [20], called earlier also view-based search [16,6], is to index data items along orthogonal category hierarchies, i.e., facets [15] (e.g., places, times, document types etc.) and use them for searching and browsing: the user selects in free order categories on facets, and the data items included in the selected categories are considered search results. After each selection, a count is computed for each category showing the number of results, if the user next makes that selection. In contrast to static filters, facets interact with each other; facets have also been called dynamic ontologies [19]. In this way, search is guided by avoiding annoying "no hits" results. Moreover, hit distributions on facets provide the end-user with data-analytic views on what kind of items there are in the underlying database. Faceted search is especially useful on the Semantic Web where hierarchical ontologies used for data annotation provide a natural basis for facets, and reasoning can be used for mapping data to facets [6].

Following facets are available in the Panel of Facets of Fig. 1: personal attributes (Family Name, First Name, Gender, Occupation, and Place of Birth), political characteristics (Chamber, Political Party, Representing State, and Serving Congress), and external datasets he or she has been linked to (Wikipedia, DBpedia, and Twitter).

People often distinguish between what data is indelible and what is ephemeral. Therefore, it is important to acknowledge that every aspect of historical entities yields information that may be of use to certain users. The faceted interface gives the user the freedom to make filtering choices freely, which provides flexibility needed in biographical and prosopographical research.

**Map Visualization View** By selecting link "Map" in the menu bar of the user interface (Fig. 1), a map visualization application depicted in Fig. 2 on the left is opened. It illustrates the intellectual mobility of Congress members through the places of their birth and death. Blue circles show birth places, red circles death places, and the diameter of the circle indicates the number of births and deaths in the place. By clicking on a circle, a pop-up window listing legislators who were born or died at the place is shown with links to their home pages in the application. This process is illustrated graphically in the right hand side of Fig. 2. The same filters and facets (on the left in Fig. 2) as in Fig. 1 can be used for filtering the target group visualized on the map. Angular Google Maps[16] is used in this visualization on top of the SPARQL endpoint.

The use case of this application view is prosopographical: the idea is to locate, map, and explain historical trends of legislator groups in the geographical space.

**Statistical Visualization View** By clicking the link "Statistics" in the menu bar of Fig. 1, two data analytic application views to the data can be opened: one based on 1) charts (pie charts and a sankey chart), and 2) one using graphs (histograms). In both cases, the target group can be filtered flexibly using filters and facets in the same way as in the previous views.

---

[15]The idea of facets dates back to the Colon Classification system of S. R. Ranganathan in library science, published in 1933.

[16]http://angular-ui.github.io/angular-google-maps/
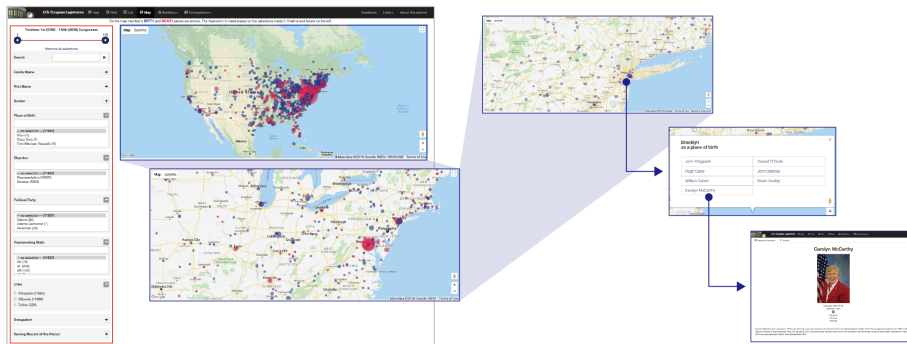
**Fig. 2.** Map Visualization View (places of birth and death of a filtered target group)

To examine the data through structured charts, and to provide glanceable overviews of the temporal features of the legislators, this page generates statistics using Google Chart diagrams[17] based on the extracted filtering results. For example, Fig. 3 LEFT illustrates the political party (Republican, Federalist, Jackson, Adams, etc.) and occupation (politician, lawyer, judge, diplomat, etc.) distributions of the legislators of the first 20 congresses, totaling 1567 members.



**Fig. 3.** Statistical Visualization View (LEFT: Pie chart, RIGHT: Histogram)

Statistics based on histograms provide yet another perspective to the data. For example, the first histogram of Fig. 3 RIGHT depicts the age distribution of the legislators

[17]https://developers.google.com/chart/

(up) and the second one about the longevity of their service (down) in years. By clicking on a bar, the links to the corresponding legislators' home pages are shown and can be inspected easily, as illustrated on the right side of the figure.

**Comparison visualization View** Finally, there is a link "Comparisons" in the menu bar that allows the user to examine the similarities and differences between the Democratic and Republican parties. In these visualizations, all functions and visualizations used in the other views (Grid, Map, Chart, and Graph), are implemented, and customized to be shown in two separate sections. The use case here is to identify and compare the properties of the two different target groups.

For example, all four comparison views are depicted in Fig. 4 illustrating the differences between the members of the Democratic and Republican parties without any further filter and facet selections made.



| Grid (Individuals) | Map | Chart | Graph |
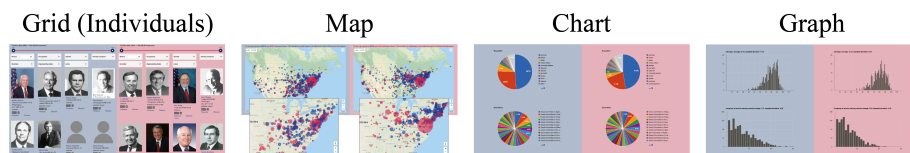
**Fig. 4.** Comparison Visualization View (Four Comparison Views)

**Use Case Examples** The Comparison visualizations can be used in different research studies. For example, it can be shown that during the Reconstruction era from the 38th through the 45th Congresses (1863–64 to 1877–78) there is a large difference in the locations of birth and death of the legislators. Most legislators were born and died in the eastern side. However, the distribution reveals a further clear tendency during this period: while the Democrats have a longitudinal spreading, Republicans remain in the Northeastern megalopolises. (cf. Fig. 5)

Another example is from the 84th through the 89th Congresses (1955–56 to 1965–66) when the federal government aimed to revitalize cities though funding urban renewal programs.[18] During this period of time, the poor were displaced and suffered from the series of policies. However, comparing with the overall trend in longevity of service of legislators which continuously decreases (cf. Fig. 6 LEFT), there is a wide variation in the longevity during 1955–1966 (cf. Fig. 6 RIGHT). This indicates that incumbent re-election rates were extremely high in both parties during this time, despite the fact that the social situation was very unstable.

It is also possible to narrow down the target groups (through filtering) and investigate the changes in a longer time period. For instance, in the sequence of decades from the 97th through the 115th Congresses (1981–82 to 2017–18), the end user is able to examine the changes in women's occupations over the years. In Fig. 7, the pie charts show the change of occupations in every ten years. Although the number of women is

---

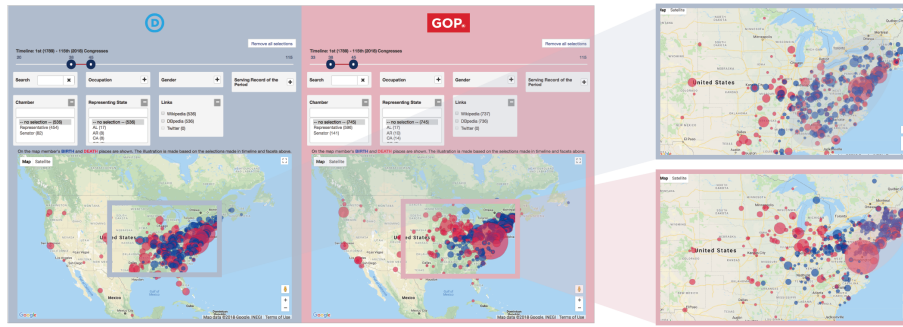[18]Widely known as "The Urban Renewal Projects"

**Fig. 5.** Birth and Death Places (TOP: Democratic, BOTTOM: Republican)



Non-Filtered (1st through 115th Congresses)

Filtered to 84th through 89th Congresses

**Fig. 6.** Longevity of Service in Histograms (LEFT: Non–Filtered, RIGHT: Filtered)

quite small in comparison to men, the result shows that the types of occupations are growing ever more diverse. Furthermore, by taking a closer look at the occupational categories from the 2000's, *social worker* and *nurse* are ranked high in Democrats (although *politician* and *lawyer* accounts for a large percentage in both parties). On the Republican side, the results are more dispersed, and *rancher* follows after *politician* and *lawyer* in the 2010's, which does not appear in the Democrats' occupations.

Through revealing such correlated continuities and changes, these examples demonstrate how historical patterns correspond to biographical information and further intertwine with politics, economics, and historical knowledge.

## 4 Implementation

*U.S. Congress Prosopographer* was implemented by extending SPARQL Faceter [11], a tool for creating faceted search interfaces on top of a SPARQL endpoint. AngularJS[19]

---

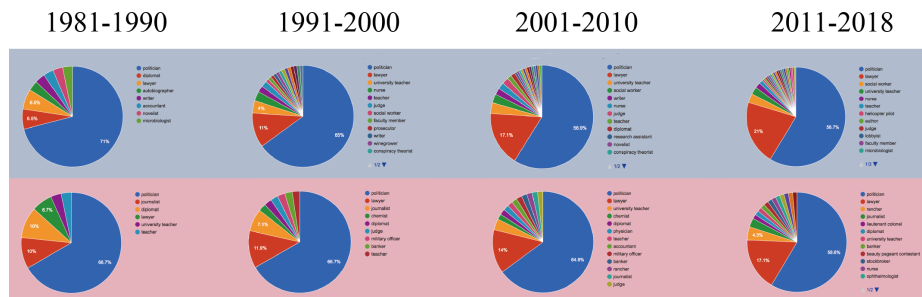[19] http://angularjs.org

**Fig. 7.** Changes in Women's Occupations (TOP: Democratic, BOTTOM: Republican)

framework was used to organize Linked Data together with a timespan slider[20] that is included as a canonical facet for the user to specify a desired range of congresses based on their interest.

Based on our experiment, SPARQL Faceter could be fairly easily combined with various means and tools for visualization with leaving a large space for customization. Modifying SPARQL Faceter itself for new features, e.g., to use functionally new kind of facets, requires expertise is using AngularJS, but just adapting the existing system to a new dataset is more straightforward. In our case, an old SPARQL Faceter application [9] was used as a template to start with.

The tool *U.S. Congress Prosopographer* contains two separate components: 1) The Linked Open Data service for machines that supports the LD publishing principles [5]. The data used in the application is available as linked open data at the "7-star" Linked Data Finland platform. 2) The user interface with application views for humans which are implemented completely on the client side as a Rich Internet Application.

Detaching the data service from applications is beneficial: the data service can be re-used by anyone for other application purposes easily without server side programming and concerns. In this way, also the computational burden of doing analyses can be distributed from the server to the clients. However, the data has to be transferred from the server to the clients for analysis, which can be problematic, when transferring and analyzing very large target groups.

## 5 Related Work and Discussion

There are a number of prior studies—in both Digital Humanities methods and tools—related to American Congress history [1,2,4,15]. However, these works have limitations in their generality and usage, restricted to particular subject, time frame, or resources (e.g., non-machine readable formats). The research of this paper stands unique in providing a comprehensive coverage of U.S. Congressional biography from its beginning until today, distributed in a Linked Open Data structure, and further enabling access

---

[20]https://github.com/angular-slider/angularjs-slider

through its "rich-prospect interface" [18] to enhance the user's understanding and exploration in American political history.

Applying Linked Data principles to cultural heritage data [7] and historical research [13] is a promising approach to solve the interoperability problems of isolated and semantically heterogeneous data sources. Linked Data has been used before also for representing and studying biographical data about the U.S. legislators [12]. Consistent with, but extending from this past research, the novelty of our work is bringing in the facet-based tooling with integrated visualizations together with a generous [22] and rich-prospect interface.

Representing and visualizing temporal data in a Linked Data format, this research allowed learners to interact with diverse data elements on both micro to macro levels. The visualizations provide both practical and comprehensive insights into the long history of the United States. On the other hand, the tool *U.S. Congress Prosopographer* explores different ways to support prosopographical research. Its different types of visualization for target groups establish context and maintain orientation while revealing details also about the individuals. This combination of macro- and microscopic viewpoints offers both qualitative and quantitative understanding of the biographical and prosopographical aspects of the Congress legislators.

This paper showed that linked open data can be used as a flexible basis for representing biographical registries, for filtering out target groups of persons of interest, and for conducting biographical and prosopograhical research. According to our practical experiences, the technology is useful, stable and handy with having the basis in Linked Data publishing standards and principles. Alongside the SPARQL querying infrastructure, the interface conveys prosopographical insights in different ways, depending on the user's curiosity and interests. In practice, assembling biographies as well as prosopographies based on Linked Data has its potential to align with traditional humanities research, not only to form research questions, but also to stimulate exploration in every individual through enriching the very foundation of historical entities within the spatial, relational, and temporal features. Once the users know and understand the latent biographical relationships as well as independent variables, they will be able to find new insights.

Like in other areas of data-driven research, the resulting interface established a robust infrastructure to explore historical records. Yet, additionally, this paper provides a new method in interpreting intangible cultural heritage resources through the dynamic integration of querying and visualizing Linked Data under one single system.

## 6  Future Work

In future work, we hope to do more experimental studies on using the tool with experts on political history of the United States. On the technical side, solutions for showing new types of visualizations, such as linguistic analyses of biographical texts or social networks of politicians, could be studied. Also, to understand broader implications and the potential influence of U.S. Congress legislators on a global scale, this research can leverage and link across other fields and domains, e.g., congressional bills, cultural enhancements, or heritage datasets of other countries. Still another direction is the aes-

thetic qualities. The visualizations are generated using standardized templates, e.g., web frameworks such as Google Charts, yet balancing between usability and design aesthetics can be studied further.

Although the research of this paper has focused on one particular dataset, the versatility of the presented general framework can be fully applied to other large scale datasets, and further give rise to new perspectives and paradigms in exploring cultural and historical spheres. Indeed, alongside with the research presented in this paper, we have transformed the National Biography of Finland, 13 000 short textual biographies of notable Finns, into a Linked Data service, and created a semantic portal for biographical and prosopographical research on top of it [10].

# References

1. Congressional biographical directory (2002), http://bioguide.congress.gov/biosearch/biosearch.asp, United States
2. Culpepper, J.: Chronicling America: Historic American Newspapers. Reference Reviews 21(7), 52–53 (2007)
3. Guha, R.V., Brickley, D., Macbeth, S.: Schema. org: evolution of structured data on the web. Communications of the ACM 59(2), 44–51 (2016)
4. Harward, B.M., Moffett, K.W.: The Calculus of Cosponsorship in the U.S. Senate. Legislative Studies Quarterly 35(1), 117–143 (2010)
5. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/
6. Hyvönen, E., Saarela, S., Viljanen, K.: Application of ontology-based techniques to view-based semantic search and browsing. In: The semantic web: research and applications. First European Semantic Web Symposium (ESWS 2004). pp. 92–106. Springer-Verlag (2004)
7. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA, USA (2012)
8. 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, Revised Selected Papers. pp. 226–230. Springer-Verlag (May 2014)
9. Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., Sirola, L.: Reassembling and enriching the life stories in printed biographical registers: Norssi high school alumni on the Semantic

---

Web. In: Language, Technology and Knowledge. First International Conference, LDK 2017, Galway, Ireland, June 19-20, 2017. Springer-Verlag (2017)

10. Hyvönen, E., Leskinen, P., Tamper, M., Tuominen, J., Keravuori, K.: Semantic National Biography of Finland. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018). pp. 372–385. CEUR Workshop Proceedings, Vol-2084 (March 2018)

11. Koho, M., Heino, E., Hyvönen, E.: SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In: Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., Vander Sande, M. (eds.) Joint Proc. of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. CEUR Workshop Proceedings, Vol-1615 (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf

12. Larson, R.: Bringing lives to light: Biography in context (2010), http://metadata.berkeley.edu/Biography_Final_Report.pdf, Final Project Report, University of Berkeley

13. Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., Van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web 6(6), 539–564 (2015)

14. Miyakita, G., Leskinen, P., Hyvönen, E.: U.s. congress prosopograher - a tool for prosopographical research of legislators (May 2018), submitted

15. Nelson, R.K., Winling, L., Marciano, R., Nathan Connolly, e.a.: American panorama (2015), https://dsl.richmond.edu/panorama/, ed. Robert K. Nelson and Edward L. Ayers

16. 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

17. Roberts, B.: Biographical Research. Understanding social research, Open University Press (2002), https://books.google.fi/books?id=04ScQgAACAAJ

18. Ruecker, S., Radzikowska, M., Sinclair, S.: Visual interface design for digital cultural heritage: A guide to rich-prospect browsing. Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing pp. 1–197 (2011)

19. Sacco, G.M.: Dynamic taxonomies: guided interactive diagnostic assistance. In: Wickramasinghe, N. (ed.) Encyclopedia of Healthcare Information Systems. Idea Group (2005)

20. Tunkelang, D.: Faceted search, Synthesis lectures on information concepts, retrieval, and services, vol. 1. Morgan & Claypool Publishers (2009)

21. Verboven, K., Carlier, M., Dumolyn, J.: A short manual to the art of prosopography. In: Prosopography Approaches and Applications. A Handbook, pp. 35–70. University of Ghent (2007), http://hdl.handle.net/1854/LU-376535

22. Whitelaw, M.: Generous interfaces for digital cultural collections. Digital Humanities Quarterly 9(1) (2015), http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html