Combining Faceted Search with Data-analytic Visualizations on Top of a SPARQL Endpoint

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

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

Abstract. This paper discusses practical experiences on creating data-analytic visualizations in a browser, on top of a SPARQL endpoint based on the results of faceted search. Four use cases related to Digital Humanities research in proposography are discussed where the SPARQL Faceter tool was used and extended in different ways. The Faceter tool allows the user to select a group of people with shared properties, e.g., people with the same place of birth, gender, profession, or employer. The filtered data can then be visualized, e.g., as column charts, with business graphics, sankey diagrams, or on a map. The use cases examine the potential of visualization as well as automated knowledge discovery in Digital Humanities research.

Keywords: Linked Data, Visualization, Biography, Prosopography, Knowledge Discovery

CEUR-WS.org/Vol-2187/paper5.pdf

1 Client-side Faceted Search on a SPARQL Endpoint

Faceted search and browsing [5,21], known also as view-based search [19] and dynamic hierarchies [20], has become a norm in web applications. The idea here is to index data items along orthogonal category hierarchies, i.e., facets [4] (e.g., places, times, document types etc.) and use them for searching and browsing: the user selects categories on facets in free order, 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 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 heterogeneous data to facets [8]. The idea of combining faceted search and visualizations has been applied, e.g., in ePistolarium[5]. However, this application is not based on Linked Data unlike ours [10,11,16,18].

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

[5] http://ckcc.huygens.knaw.nl/epistolarium/

Faceted search can be implemented with server-side solutions, such as Solr[6], Sphinx[7], and ElasticSearch[8], and higher level tools, such as vuFind[9]. However there is a lack of light-weight client-side faceted search tools or components that are able to search large datasets directly from a SPARQL endpoint. Such a tool is useful, because it can be used easily on virtually any open SPARQL endpoint on the web without any need for server side programming and access rights. This paper presents such a tool, SPARQL Faceter, a web component for implementing faceted search applications efficiently in a browser, based only on a standard SPARQL API. We extend our earlier short paper of the tool [14] by 1) showing detail about how the tool is used and works, by 2) explaining novelties in its latest version, and 3) especially show how the tool and faceted search can be extended with different kind of data-analytic visualizations.

As a proof of concept, four use case studies of data visualization are discussed from a SPARQL Faceter perspective: 1) WarSampo, using cultural heritage materials of World War II in Finland [10]. 2) Norssit, on top of a Finnish high school alumni registry data [11]. 3) Semantic National Biography of Finland, based on the National biography of the Finnish literature society [16]. 4) U.S. Congress Prosopographer, utilizing biographical records of U.S. Congress legislators [18]. In these cases, the following two-step prosopographical research method [22, p. 47] is supported where the goal is to find out some kind of commonness or average in selected *target groups* of people. 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 possibly compared with other groups, in order to solve the research question. For finding target groups, faceted search is used, and then visualizations are created in order to analyze their characteristics.

The rest of the paper is organized as follows. First, characteristics of SPARQL Faceter are explained with a focus on showing how it is used in practice in applications. After this, extending the tool with visualizations is in focus. In conclusion, lessons learned and directions for further research are discussed.

2 Using and Extending SPARQL Faceter

SPARQL Faceter uses AngularJS[10] as the implementation framework. The GitHub page[11] gives instructions how to install it, and how to define the application with facets of desired type in the source code. The page provides demo examples with queries to DBpedia and WarSampo databases. The developer can adopt it to any Linked Data publication by configuring the endpoint, property paths for facets, and queries. The SPARQL Faceter is documented in detail[12].

[6] http://lucene.apache.org/solr/
[7] http://sphinxsearch.com/blog/2013/06/21/faceted-search-with-sphinx/
[8] https://www.elastic.co/
[9] https://vufind.org/
[10] https://angularjs.org/
[11] https://github.com/SemanticComputing/angular-semantic-faceted-search
[12] http://semanticcomputing.github.io/angular-semantic-faceted-search/#/api

```
PREFIX xsd:     <http://www.w3.org/2001/XMLSchema#>
PREFIX schema:  <http://schema.org/>
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl:  <http://www.w3.org/2008/05/skos-xl#>
PREFIX nbf:     <http://ldf.fi/nbf/>
PREFIX crm:     <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX foaf:    <http://xmlns.com/foaf/0.1/>
PREFIX gvp:     <http://vocab.getty.edu/ontology#>

SELECT DISTINCT (?id AS ?id__uri) ?id__name ?value WHERE {
  # Restraints set in Faceter
  { ?id a nbf:PersonConcept ;
        foaf:focus/^crm:P98_brought_into_life/
        nbf:time/gvp:estStart ?slider_2 .
    FILTER (1800<=year(?slider_2) && year(?slider_2)<=2018)
    }

  # Query person's age
  ?id foaf:focus/^crm:P100_was_death_of/nbf:time
      [ gvp:estStart ?time ; gvp:estEnd ?time2 ] ;
      foaf:focus/^crm:P98_brought_into_life/nbf:time
      [ gvp:estStart ?birth ; gvp:estEnd ?birth2 ] .
  BIND (xsd:integer(0.5*
      (year(?time)+year(?time2)-year(?birth)-year(?birth2)))
    AS ?value)
  # Filter out erroneous cases
  FILTER (-1<?value && ?value<120)

  # Query for person's name
  ?id skosxl:prefLabel ?id__label .
  OPTIONAL { ?id__label schema:familyName ?id__fname }
  OPTIONAL { ?id__label schema:givenName ?id__gname }
  BIND (CONCAT(
      COALESCE(?id__gname, ""),
      " ",
      COALESCE(?id__fname, ""))
    AS ?id__name)

} ORDER BY ?value ?id__fname ?id__gname
```

Fig. 1. A SPARQL example for querying people's ages

Figure 1 depicts a SPARQL query for fetching the data visualized in Fig. 2. The first code block in the query pattern defines the restricted target group of the Faceter application, in this case we are interested in people who were born on or after the year 1800, a choice that has been made with the timespan slider. The example follows the data model of the National Biography of Finland [16], so to query for the desired resource in the data property paths are utilized. In the next block related events of birth are searched, and the age of a person is calculated. Errors in the data are filtered out by accepting only values in the range of 0 to 120 years. In the third block person's proper name is constructed. Some of the fields are optional because we cannot assume all person entries to have both a first and a family name. The query returns a JSON formatted array consisting of objects containing the URI of the resource, the person name, and age. Notice that for linkage purposes, we also need the person information, not as histogram data with ages as bins with corresponding counts. In the application

the data is converted to a JavaScript array suitable for Google Chart tools[13]. The output in this example case, (Fig. 2) is a column chart with age on the horizontal, and the amount of people on the vertical axis. A mouse click on any of the columns opens a modal list of all people having that age, from which the user can get to the detailed page of a person.

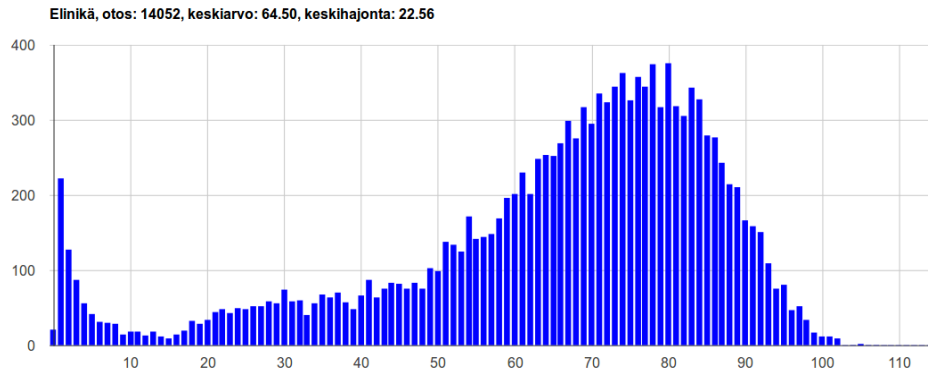

Fig. 2. Lifespan of people lived in 1800–2018.

3 Applications

In this section examples of visualizations on top of the SPARQL Faceter tool in different applications are shown and discussed.

1) WarSampo is the first semantic portal for serving and publishing large heterogeneous sets of linked open data about the World War II (WW2)[14]. To create a global view of the war, and to attain a deeper understanding about its history, the portal contains e.g., some 95 000 death records of the Finnish WW2 casualties. This in-use portal includes 8 different application perspectives through different datasets, and had 130 000 users in 2017.

Fig. 3 shows a screenshot of the faceted search application in the casualties perspective[15]. The data is laid out in a table-like view. Facets are presented on the left of the interface with string search support. The number of hits on each facet is calculated dynamically and shown to the user, and selections leading to an empty result set are hidden. In Fig. 3, nine facets and the results are shown, where the user has selected "widow" in the marital status facet, limiting the search down to 279 killed widows that are presented in the table with links to further information.

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

[14] https://www.sotasampo.fi/en/

[15] https://www.sotasampo.fi/en/casualties/

The faceted search is used not only for searching but also as a flexible tool for researching the underlying data. In Fig. 3, the hit counts immediately show distributions of the killed widows along the facet categories. For example, the facet "Number of children", which is not visible in the figure, shows that one of the deceased widows had 10 children and most often (in 89 cases) widows had one child. If we next select the category "one child" on its facet, we can see that two of the deceased are women and 86 are men in the gender facet. In the latest version of SPARQL Faceter, each facet component has a push button for visualizing the distributions with Google pie charts.

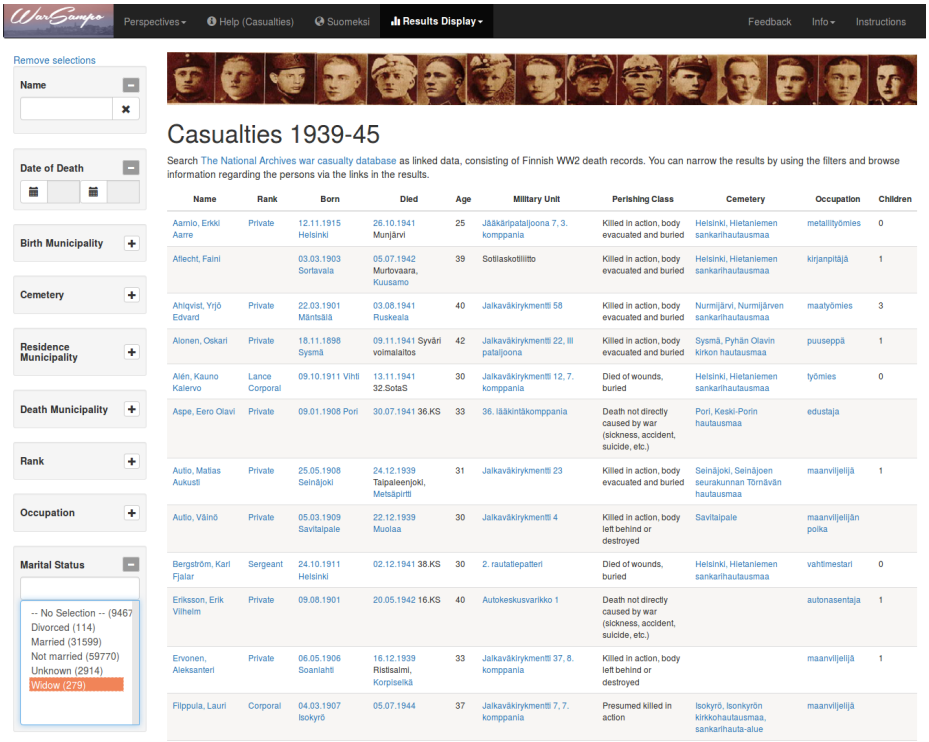

Fig. 3. The faceted search interface of death records with one selected facet. The left side contains the facets, displaying available categories and the amount of death records for each selection. Death records matching the current facet selections are shown as a table.

In addition to the table-like results view, the casualties perspective allows to visualize the results using three different visualization types: 1) age distribution (column chart), 2) personal life paths (sankey diagram), and 3) distribution of property values (bar chart) [13]. The method of results display does not affect the operation of facets, so the user can use same facet selections, and have different views of the resulting death records. Figure 4 shows a visualization of the life paths of soldiers that were buried in the cemetery of the municipality of Kolari. The visualization shows the known munic-

ipalities where the individuals have been at different times of their lives, which also reveals the fact that perished Finnish soldiers were usually buried in their hometown.

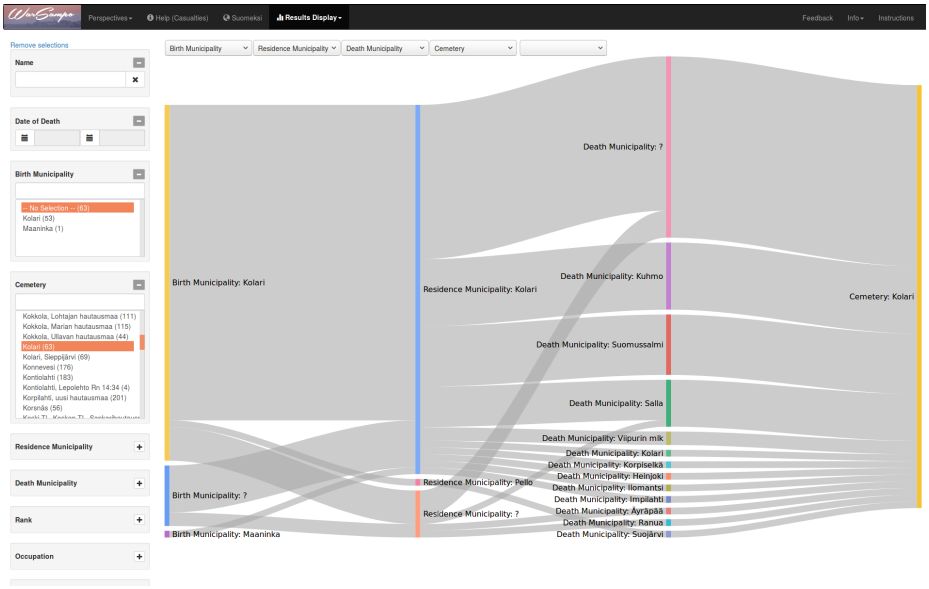

Fig. 4. Life paths of 63 soldiers buried in the cemetery of Kolari.

In addition to the casualties perspective, the WarSampo portal employs SPARQL Faceter in two other perspectives, photographs and cemeteries, to provide a user interface to search, browse and explore their data content.

2) Norssit dataset consists of a register with over 10 000 alumni of the prominent Finnish high school "Norssi" in 1867–1992. The register was transformed into RDF, was enriched by data linking, was published as a linked data service, and is provided to end users via a faceted search engine and browser for studying their lives and for prosopographical research. [11]

The Norssit portal[16] contains two pages for visualizations[17,18]. The pages use Google Charts showing search results as pie charts, column charts, or sankey diagrams [15]. An example of rendering the most common employers on different decades is depicted in Fig. 5.

3) Semantic National Biography of Finland The National Biography of Finland[19] consists of biographies of notable Finnish people throughout history. The biographies

[16] http://www.norssit.fi/semweb

[17] http://www.norssit.fi/semweb/#!/visualisointi

[18] http://www.norssit.fi/semweb/#!/visualisointi2

[19] http://kansallisbiografia.fi

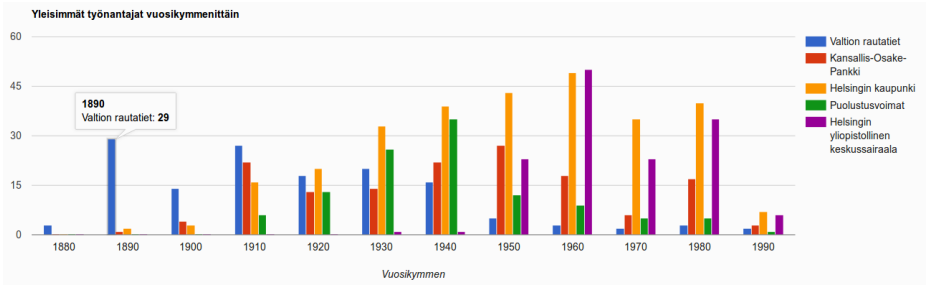

Fig. 5. Column chart showing the most common employers and their changes in time.

describe the lives and achievements of historical figures, containing vast amounts of references to notable Finnish and foreign figures, including internal links to other biographies. [12]

To support the prosopographical research, the portal contains pages with faceted search where the data is visualized on Google Maps[20], or as column charts [15]. An example of rendering the query results on a map is depicted in Fig. 6. The portal also has a faceted search page for linguistic analysis of the vocabulary used in biographical descriptions.

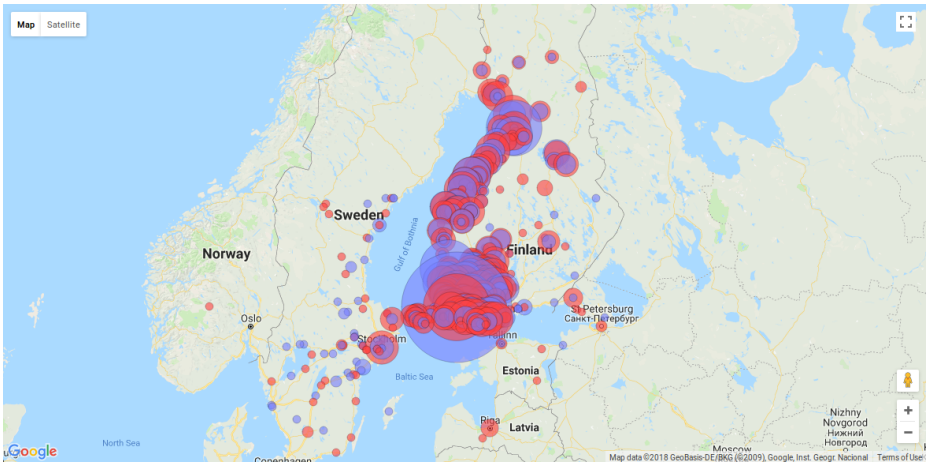

Fig. 6. Places of birth and death of 17th century Finnish clergy

[20] https://cloud.google.com/maps-platform/

4) U.S. Congress Prosopographer This interface[21] contains biographical records of 11 987 people who served in the U.S. Congresses from the 1st (1789) to the 115th (2018) one—converted and extracted from open-source data[22,23]. The interface contains four integrated tools and demonstrates how historical patterns correspond to biographical information and further intertwine with politics, economics, and historical knowledge alongside the American history.

Being adapted from the previous studies above, novelty of this interface are the comparing visualizations. As shown in Fig. 7, a different set of target groups—in this case, the two major parties, Democrats and Republicans—can be analyzed and compared with each-other. The end user is able to find and execute new insights through the independent variables, as well as the latent biographical relationship of U.S. Congress legislators through selecting, filtering, and comparing two different accounts of histories.

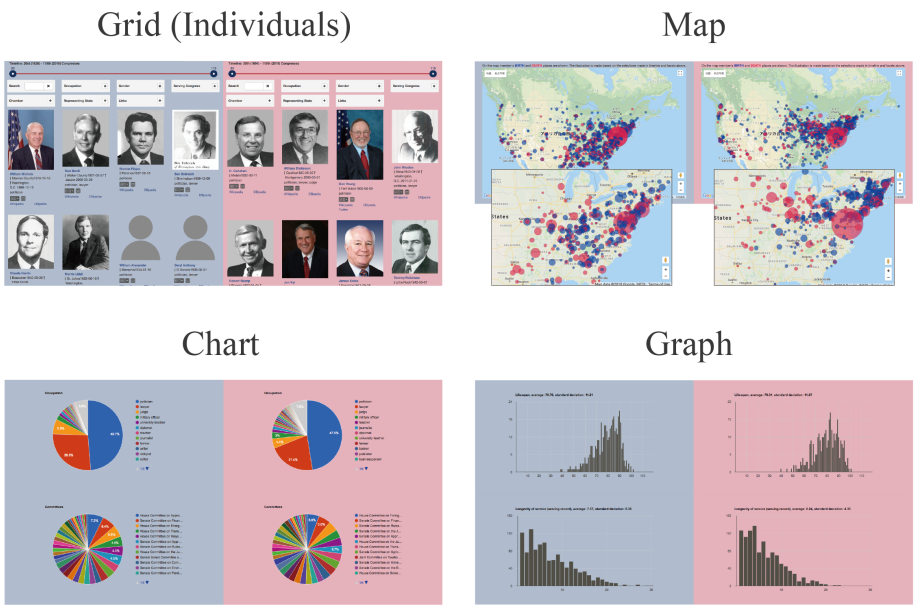

Fig. 7. Examples of Comparing Visualizations: Democratic (left) / Republican (right)

[21] https://semanticcomputing.github.io/congress-legislators/
[22] https://github.com/unitedstates/congress-legislators
[23] http://k7moa.com

4 Related Work and Discussion

The conference proceedings [2] include several papers on bringing biographical data online, on analyzing biographies with computational methods, on group portraits and networks, and on visualizations. Applying Linked Data principles to cultural heritage data [9] and historical research [17] has been a promising approach to solve the problems of isolated and semantically heterogeneous data sources. Also a number of previous research exists in faceted search [6,7] and Linked Data visualization [1,3,4].

Based on the applications discussed, faceted search and browsing can be combined in a useful way with various means and tools for visualization: facet selections are a very flexible way to filter out result sets, and we have demonstrated that this can be done in real time using SPARQL queries in endpoints containing tens of millions of triples. Based on the query results, wrappers for data visualization tools, such as Google Charts for statistics or network analysis tools can be integrated and reused easily. By making the data analysis on the client side, computational burden can be distributed to end-user browsers, and Rich Internet Applications can be created without server-side programming. Moreover, the resulting visualizations open up ways of exploring new types of questions, and further evoke a knowledge discovery process in conducting digital humanities research.

A key challenge in this approach is how to deal with large result sets. It is usually not feasible to transfer very large result sets, say tens of thousands of casualty records in the WarSampo case, from the server to the browser. If the data is not available in the browser, it cannot of course be analyzed there. This problem is solved in SPARQL Faceter by paginating the results; the results are uploaded in pages and only when needed. The end-user should be aware about the limitation that the visualizations are based on only the data that has been uploaded. The size of the page therefore sets a limit on how large datasets can be visualized, even though very large result datasets can be queried on the server side.

The use case study WarSampo was implemented by the original SPARQL Faceter [14] while the other use cases discussed are based on its new versions with the following enhancements: 1) Every facet is now able to make its own SPARQL query (or many), which leads to better efficiency. 2) Hierarchical facets up to any number of levels are supported and more efficiently implemented. 3) Text search facet is included as a new facet type. 4) A slider facet for selecting a range of numerical values interactively can be used. 5) Facet hit distributions can be visualized using pie charts in addition to hit counts. There are also some enhancements made for visualizations. The U.S. Congress Prosopographer and Semantic National Biography allow, for example, visual comparison of two groups. The different Faceter versions and extensions need to be amalgamated together in next versions of the tool and applications. Also the technical solutions for showing new types of visualization, such as social networks of people, should be studied more—we are about to implement a network application to the Semantic National Biography. Still another direction for further work are the aesthetic qualities. The visualizations are generated using standardized templates, e.g., web frameworks such as Google Charts, and balancing between usability and design aesthetics needs to be studied.

Acknowledgements Thanks to Erkki Heino for implementational help regarding extending the Faceter SPARQL tool for our case studies, and to Jouni Tuominen for discussions related to the data models and data services underlying our applications. Goki Miyakita was supported by a mobility scholarship at Aalto University in the frame of the Erasmus Mundus Action 2 Project TEAM Technologies for Information and Communication Technologies, funded by the European Commission. Our research was also supported by the CSC computing services and the Severi project[24] funded mainly by Business Finland.

References

1. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: A survey of the state of the art. In: Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference. CEUR Workshop Proceedings, Vol-1558 (2016)
2. ter Braake, S., Anstke Fokkens, R.S., Declerck, T., Wandl-Vogt, E. (eds.): BD2015 Biographical Data in a Digital World 2015. CEUR Workshop Proceedings, Vol-1399 (2015), http://ceur-ws.org/Vol-1399/
3. Dadzie, A.S., Pietriga, E.: Visualisation of linked data–reprise. Semantic Web 8(1), 1–21 (2017)
4. Dadzie, A.S., Rowe, M.: Approaches to visualising Linked Data: A survey. Semantic Web 2(2), 89–124 (2011)
5. 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)
6. Heim, P., Ziegler, J., Lohmann, S.: gFacet: A browser for the web of data. In: Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW 2008). CEUR-WS, vol. 417, pp. 49–58 (2008), http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-417/paper5.pdf
7. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: A browser for heterogeneous semantic web repositories. In: International Semantic Web Conference. pp. 272–285. Springer (2006)
8. 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)
9. 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)
10. Hyvönen, E., Heino, E., Leskinen, P., Ikkala, E., Koho, M., Tamper, M., Tuominen, J., Mäkelä, E.: WarSampo data service and semantic portal for publishing linked open data about the second world war history. In: The Semantic Web – Latest Advances and New Domains (ESWC 2016). pp. 758–773. Springer-Verlag (2016)
11. 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)
12. 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)

[24] http://seco.cs.aalto.fi/projects/severi

13. Ikkala, E., Koho, M., Heino, E., Leskinen, P., Hyvönen, E., Ahoranta, T.: Prosopographical views to finnish ww2 casualties through cemeteries and linked open data. In: Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). CEUR Workshop Proceedings (October 2017)
14. 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. No. 1615, CEUR Workshop Proceedings (2016), http://ceur-ws.org/Vol-1615/semdevPaper5.pdf
15. Leskinen, P., Hyvönen, E., Tuominen, J.: Analyzing and visualizing prosopographical linked data based on short biographies. In: BD2017 Biographical Data in a Digital World 2017, Proceedings. CEUR Workshop Proceedings (2018)
16. Leskinen, P., Tuominen, J., Heino, E., Hyvönen, E.: An ontology and data infrastructure for publishing and using biographical linked data. In: Proceedings of the Workshop on Humanities in the Semantic Web (WHiSe II). pp. 15–26. CEUR Workshop Proceedings, Vol-2014 (2017)
17. 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)
18. Miyakita, G., Leskinen, P., Hyvönen, E.: U.S. Congress Prosopographer – A Tool for Prosopographical Research of Legislators (May 2018), submitted
19. 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
20. Sacco, G.M.: Dynamic taxonomies: guided interactive diagnostic assistance. In: Wickramasinghe, N. (ed.) Encyclopedia of Healthcare Information Systems. Idea Group (2005)
21. Tunkelang, D.: Faceted search, Synthesis lectures on information concepts, retrieval, and services, vol. 1. Morgan & Claypool Publishers (2009)
22. 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