Life Stories as Event-based Linked Data: Case Semantic National Biography

Eero Hyvönen, Miika Alonen, Esko Ikkala, and Eetu Mäkelä

Semantic Computing Research Group (SeCo), Aalto University
http://www.seco.tkk.fi/, firstname.lastname@aalto.fi

Abstract. This paper argues, by presenting a case study and a demonstration on the web, that biographies make a promising application case of Linked Data: the reading experience can be enhanced by enriching the biographies with additional life time events, by proving the user with a spatio-temporal context for reading, and by linking the text to additional contents in related datasets.

1 Introduction

This paper addresses the research question: *How can the reading experience of biographies be enhanced using web technologies?* Our research hypotheses is to apply the Linked Data (LD) approach to this, with the idea of providing the reader with a richer reading context than the biography document alone. The focus of research is on: 1) *Data linking*. Biographies can be linked with additional contextual data, such as links to the literal works of the person. 2) *Data enriching*. Data from different sources can be used for enriching the life story with additional events and data, e.g., with metadata about a historical event that the person participated in. 3) *Visualization*. LD can be visualized in useful ways. The life story can, e.g., be shown on maps and timelines. We tested the hypoheses in a case study[1] where the Finnish National Biography[2] (NB), a collection of 6,381 short biographies, is published as LD in a SPARQL endpoint with a demonstrational application based on its standard API.

2 Representing Biographies as Linked Data

To enrich and link biographical data with related datasets the data must be made semantically interoperable, either by data alignments (using, e.g., Dublin Core and the dumb down priciple) or by data transformations into a harmonized form [3]. In our case study we selected the data harminization approach and the event-centric CIDOC CRM[3] ISO standard as the ontological basis, since biographies are based on life events. NB biograhies are modeled as collections of CIDOC CRM events, where each event is characterized by the 1) actors involved, 2) place, 3) time, and 4) the event type.

[1] Our work was funded by Tekes, Finnish Cultural Foundation, and the Linked Data Finland consortium of 20 organizations.

[2] http://www.kansallisbiografia.fi/english/?p=2

[3] http://www.cidoc-crm.org/

A simple custom event extractor was created for transforming biographies into this model represented in RDF. The extractor first lemmatizes a biography and then analyzes its major parts: a textual story followed by systematically titled sections listing major achievements of the person, such as “works”, “awards”, and “memberships” as snippets. A snippet represents an event and typically contains mentions of years and places. For example, the biography of architect Alvar Aalto tells “WORKS: ...; Church of Muurame 1926-1929;...” indicating an artistic creation event. The named entity recognition tool of the Machinese[4] NLP library is used for finding place names in the snippets, and Geonames is used for geocoding. Timespans of snippet events are found easily as numeric years or their intervals, and an actor of the events is the subject person of the biography. The result of processing a biography is a list of spatio-temporal CIDOC CRM events with short titles (snippet texts) related to the corresponding person. At the moment, the extractor uses only the snippets for event creation—more generic event extraction from the free biography narrative remains a topic of further research.

For a domain ontology, we reused the Finnish History Ontology HISTO by transforming it into CIDOC CRM. The new HISTO version contains 1,173 major historical events (E5_Event in CIDOC CRM) covering over 1000 years of Finnish history, and includes 80,085 activities (E7_Activity) of different kinds, such as armistice, election etc. Linked to these are 7,302 persons (E21_Person) and a few hundred organizations and groups, 3,290 places (E53_Place), and 11,141 time spans (E52_Time-span). The data originates from the Agricola timeline[5] created by Finnish historians.

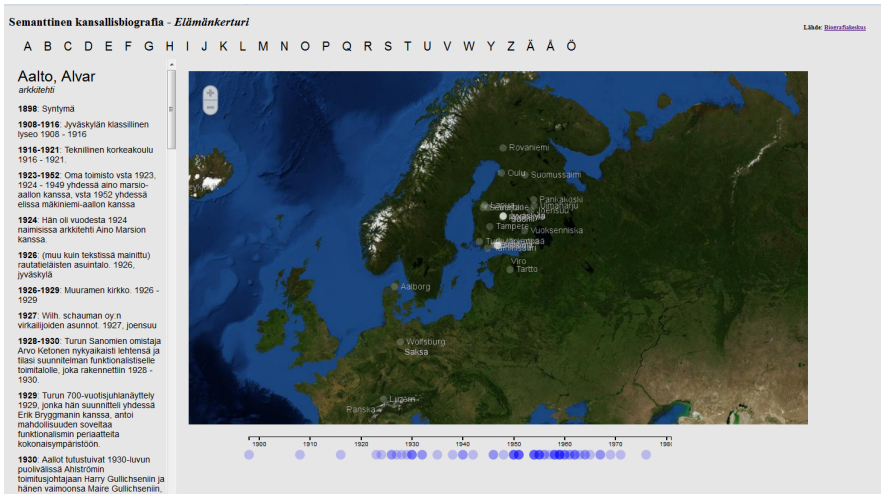

Fig. 1. Spatio-temporal visualization of Alvar Aalto’s life with external links.

[4] http://www.connexor.com/nlplib/

[5] http://agricola.utu.fi/

The extracted events were then enriched with events from external datasets as follows: 1) Persons in SNB and HISTO were mapped onto each other based on their names. This worked well without further semantic disambiguation since few different persons had similar names. NB and HISTO shared 921 persons p, and the biography of each p could therefore be enriched with all HISTO events that p was involved in. 2) There were 361 artistic creation events (e.g., publishing a book) of NB persons that could be extracted from Europeana Linked Open Data[6] using the person as the creator. Related biographies could therefore be enriched with events pointing to Europeana contents. 3) The NB persons were involved in 263 instances of publications of the Project Gutenberg data[7]. Corresponding events could therefore be added into the biographies, and links to the original digitized publications be provided. 4) The NB persons were also linked to Wikipedia for additional information; again simple string matching produced good results. These examples demostrate how linked events can be extracted from other datasets and be used for enriching other biographical events. In the experiment, 116,278 spatio-temporal events were finally extracted for the NB biography records.

3 Biographies Enriched in a Spatio-temporal Context

Based on the enriched and linked biography data, a demonstrator was created proving the end user with a spatio-temporal context for reading NB biographical data as well as links to addtional content from related sources. Fig. 1 depicts the user interface online[8] with architect Alvar Aalto's biography selected; the other 6,400 celebrities can be selected from the alphabetical list above. On the left column, temporal events extracted from the biography and related datasets are presented (in Finnish), such as "1898 Birth", and "1908-1916 Jyväskylä Classical Lyceum". The event "1930–1939: Alvar Aalto created his famous functionalist works (*Histo*)" shows an external link to HISTO for additional information. The events are also seen as bubbles on a timeline at the bottom. The map in the middle shows the end-user the places related to the biography events. By hovering the mouse over an event or its bubble the related event is high-lighted and the map zoomed and centered around the place related to the event. In this way the user can quickly get an overview about the spatio-temporal context of Alvar Aalto's life, and get links to additional sources of information. The actual biography text can be read by clicking a link lower in the interface (not visible in the figure). The user interface also performs dynamic SPARQL querying for additional external links. In our demonstration, the BookSampo dataset and SPARQL endpoint [6] is used for enriching literature-related biographies with additional publication and literature award events.

The user interface for spatio-temporal lifeline visualization was implemented using AngularJS[9] and D3[10] on top of the Linked Data Finland (LDF) data service[11].

[6] http://pro.europeana.eu/linked-open-data
[7] http://datahub.io/dataset/gutenberg
[8] http://www.ldf.fi/dataset/history/map.html
[9] http://angularjs.org
[10] http://d3js.org
[11] Cf. http://www.ldf.fi/dataset/history/ for dataset documentation and SPARQL endpoint

4 Discussion, Related Work, and Future Research

Our case study suggests that biography publication is a promising application case for LD. The event-based modeling approach was deemed useful and handy, after learning basics of the fairly complex CIDOC CRM model. The snippet events could be extracted and aligned with related places, times, and actors fairly accurately using simple string-based techniques. However, the results of event extraction and entity linking have not been evaluated formally, and it is obvious that problems grow with larger datasets and when analysing free text—these issues are a topic of future research.

Biographical data has been studied by genealogists (e.g., (Event) GEDCOM[12]), CH organizations (e.g., the Getty ULAN[13]), and semantic web researchers (e.g., BIO ontology[14]). Semantic web event models include, e.g., Event Ontology [8], LODE ontology[15], SEM [1], and Event-Model-F[16] [9]. A history ontology with map visualizations is presented in [7], and an ontology of historical events in [4]. Visualization using historical timelines is discussed, e.g., in [5], and event extraction reviewed in [2].

References

1. van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G.: Design and use of the simple event model (SEM). Web Semantics: Science, Services and Agents on the World Wide Web 9(2), 128–136 (2011)
2. Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F.: An overview of event extraction from text. In: DeRiVE 2011, Detection, Representation, and Exploitation of Events in the Semantic Web (2011), http://ceur-ws.org/Vol-779/
3. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Morgan & Claypool, Palo Alto, CA (2012)
4. Hyvönen, E., Alm, O., Kuittinen, H.: Using an ontology of historical events in semantic portals for cultural heritage. In: Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007) (2007), http://www.cs.vu.nl/ laroyo/CH-SW.html
5. Jensen, M.: Vizualising complex semantic timelines. NewsBlip Research Papers, Report NBTR2003-001 (2003), http://www.newsblip.com/tr/
6. Mäkelä, E., Ruotsalo, T., Hyvönen: How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo. Semantic Web – Interoperability, Usability, Applicability 3(1) (2012)
7. Nagypal, G., Deswarte, R., Oosthoek, J.: Applying the semantic web: The VICODI experience in creating visual contextualization for history. Lit Linguist Computing 20(3), 327–349 (2005), http://dx.doi.org/10.1093/llc/fqi037
8. Raimond, Y., Abdallah, S.: The event ontology (2007), http://motools.sourceforge.net/event/event.html
9. Scherp, A., Saathoff, C., Franz, T.: Event-Model-F (2010), http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/ontologies/events

[12] http://en.wikipedia.org/wiki/GEDCOM
[13] http://www.getty.edu/research/tools/vocabularies/ulan/
[14] http://vocab.org/bio/0.1/.html
[15] http://linkedevents.org/ontology/
[16] http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/ontologies/events