EMLO Prosopographical Data Model:
Towards a Biographical Conceptual Reference Model

STSM report, Cost Action IS1310, Reassembling the Republic of Letters
Jouni Tuominen, Aalto University, 28th April 2016

Introduction and Background of the STSM

This report is based on the COST short term scientific mission (STSM) that took place in the University of Oxford from 29th March to 8th April 2016. The purpose of the STSM was to develop a data model for representing biographical information for prosopographical research. The starting point for this work was the prosopographical data model designed for the Early Modern Letters Online (EMLO) database [1], as part of the Cultures of Knowledge (CofK) project in the University of Oxford. In addition, the contents of the EMLO prosopographical database had already been published as a Linked Data pilot as part of the CofK project [2].

The work started with analyzing the current EMLO data model and its Linked Data publication. The EMLO data model is an event-based, person- and role-centric model for representing the activities a person has participated in during his life, essentially forming his biography. A such activity involves one or more participants, acting in specific roles, place(s), and a time expression. The EMLO data model is based on PROV [3] for representing the activities and the participant roles, while the Linked Data pilot continues the use of PROV for the roles, and also introduces the CIDOC CRM (Conceptual Reference Model) [4] for representing the activities as events. Further, related data models for representing (biographical) events were inspected and compared, including: BIO [5], Relationship [6], BIBO [7], ORG [8], Event Ontology [9], LODE [10], PROV, CIDOC CRM, VIVO [11], Simple Event Model (SEM) [12], ULAN [13], and PROSO [14].

Design Principles and Features of the Developed Data Model

The design of the data model developed during the STSM is based on the anticipated use cases for prosopographical research (Robin Buning's presentation in the WG2 workshop, [15]), in the sense that the data model should support the principal tasks relevant for the researchers in this area. The development of the data model was done in cooperation with prof. Eero Hyvönen, who was visiting the University of Oxford at the same time as my STSM took place.

The outcome of the work is a data model with the following features:

- Act as an extension of CIDOC CRM for compatibility with other cultural heritage datasets.
- Separation of the general biographical data model from the EMLO-specific event types and participant roles, to support also other kinds of prosopographical datasets, concerning different cultures and time periods.

- Support for principal query types for prosopographical research: finding a set of people who share selected characteristics, and extracting networks of people based on some criterion for further analysis, e.g., with external visualization toolkits.
- Distinguish between unary roles (e.g., professions), binary relationships (e.g., family relations), and events (e.g., baptism) of person's biography for intuitive information representation and query writing, with a shared role-centric modeling approach.
- Represent the roles of participants of events as instances of OWL classes, enabling reasoning and validation of the integrity of prosopographical data created based on the model, ensuring the quality of the data (e.g., a baptism may involve only participants in the roles of baptismal candidate, officiant, etc.).

The resulting model, Biographical Conceptual Reference Model (Bio CRM), facilitates the integration of different prosopographical datasets represented with their own classification schemes for events, actors, roles, etc., in the spirit of CIDOC CRM, which aims to harmonize datasets on the field of cultural heritage.

Bio CRM introduces the following changes when compared to the current EMLO data model and its Linked Data pilot publication:

- More extensive utilization of CIDOC CRM.
- Roles are represented in the spirit of Basic Formal Ontology (BFO) instead of PROV, as PROV is meant for representing provenance information involved in producing a piece of data or thing, and all biographical events are not such activities.
- Unary roles and binary relationships are easier to represent, because they are not modeled as or dependent on events; rather events can be used to qualify them (e.g., adding temporal or spatial context). Only the role of the "target" person needs to be explicitly stated in binary relationships (e.g., "person X has a family relation to Aunt Y" vs. in EMLO: "Nephew X is in a family relationship activity with Aunt Y").
- Roles are semantically tied to the activity types they can be used in, using OWL constructs in the RDF schema of the data model. Previously the connection between the roles and activity types has been made explicit only in the Excel sheet for inputting prosopographical data.

Data Model Documentation and Dissemination

Documentation of the developed Bio CRM model including more detailed explanation of the design principles, examples and a comparison of different approaches for representing the roles of actors will be published in separate document(s) and will be shared to the COST action participants, accompanying the concrete RDF schema of the data model. There is also a plan to write a scientific article that describes the data model and rationalizes the selected modeling approach. Future work includes testing the model with real data, by converting the EMLO Linked Data pilot publication to follow the model.

The STSM also involved discussions and planning with the CofK project team on how the tools developed at Aalto University could be used to support the production of epistolary and prosopographical data in EMLO, for mapping new spreadsheet data to EMLO people and places, linking them to external ontologies, and resolving duplicate data.

References

1. Gray Jones, Tanya: Cultures of Knowledge : Prosopographical Data, 31st March 2015, https://docs.google.com/document/d/13KFFcS6MJ_c6AmOJF1TvGzVCQyuCTEvWAgkBj0RNE_I/edit?usp=sharing
2. EMLO database published as a Linked Data pilot, http://ldf.fi/emlo
3. PROV-O: The PROV Ontology, https://www.w3.org/TR/2013/REC-prov-o-20130430/
4. The CIDOC Conceptual Reference Model (CRM), http://www.cidoc-crm.org/
5. BIO: A vocabulary for biographical information, http://vocab.org/bio/
6. RELATIONSHIP: A vocabulary for describing relationships between people, http://vocab.org/relationship/
7. Bibliographic Ontology Specification, http://bibliontology.com/
8. The Organization Ontology, https://www.w3.org/TR/vocab-org/
9. The Event Ontology, http://motools.sourceforge.net/event/event.122.html
10. LODE: An ontology for Linking Open Descriptions of Events, http://linkedevents.org/ontology/
11. VIVO-ISF Ontology. https://wiki.duraspace.org/display/VIVO/VIVO-ISF+Ontology
12. The Simple Event Model Ontology, http://semanticweb.cs.vu.nl/2009/11/sem/
13. Getty Vocabularies: Linked Open Data, Semantic Representation, ULAN Specifics, http://vocab.getty.edu/doc/#ULAN_Specifics
14. Zingoni, Jacopo: PROSO data model - a solution for modelling historical academic prosopographical records as linked data through an event based ontological approach, Atelier Heloïse Workshop, 2014, http://amsacta.unibo.it/3992/
15. Buning, Robin: Prosopographical Research Questions, WG2 workshop, 22th January 2016, http://seco.cs.aalto.fi/events/2016/2016-01-22-oxford/slides/2016-01-22-Buning-Oxford-WG2.pdf