IPIF - pragmatic modelling decisions

Matthias Schlögl*[1], Georg Vogeler*[2], and Gunter Vasold*[2]

[1]Austrian Centre for Digital Humanities and Cultural Heritage (Austrian Academy of Sciences) – Austria

[2]University of Graz – Austria

Abstract

Prosopography is one of the core fields that triggered the use of digital methods in historical research. Prosopographical data collections have been converted into databases since the 1970s (e.g. Barman et al. 1977, Althoff 1978). Social network analysis has triggered a rich corpus of research on prosopographical data (e.g. Warren et al. 2016, Jackson 2017, Bühlmann 2012). It is no surprise that a discussion on modelling took up, in which John Bradley labeled an idea shared by several prosopographical researchers as the "factoid" approach (Bradley/Short 2005). The factoid approach keeps the information gathered from the sources on a person separate from its justification in a source and the abstract concept of the person.
There are two major approaches to allow data exchange: Exposing the data as RDF or via a local RESTful API. Both have advantages and disadvantages we have discussed in 2019 (Vogeler, Schlögl, Vasold, forthcoming). Following the example of IIIF, IPIF takes a pragmatic approach to data exchange: It defines a RESTful API intended to be easily consumable. It is described following the OpenAPI standard which facilitates code creation on server and client side. IPIF has been realized in several prototypes (Vogeler/Schlögl/Vasold/Stoff 2020)

The fundamental modeling decision in IPIF is to follow the factoid model: in IPIF the factoid with the metadata of its creation and modification aggregates information on a person, the source of information and the statements extracted from the source by the creator of the factoid. Most of the data can have either a literal content (label) or reference alternatively a URI (uri): places, groups, related persons, or roles. To remain open to case specific information, statements can be typed by the data provider. IPIF strives to be part of the semantic web by describing the JSON results of the API as json-ld.

While developing the API and implementing prototypes (Vogeler et al. 2020) several issues came up. We think they are common to modelling person related information from historical data sets and would like to discuss them in the present paper.

It is well known, how complicated historical dates can be: time ranges (e.g. "in summer 1450"), relative time (e.g. "before Eastern 1130"), incomplete information (e.g. by damages in the source) or information reduced just for communication (e.g. reigns expressed in years). IPIF currently does not model all these possibilities but distinguishes between date/label and date/sortdate, which allows to keep the original information in textual form but also allows to filter and sort by a computable date, which is based on the creators decision.

*Speaker

The factoid model creates a network of data resources: factoids, persons, statements, sources. All of them are represented with endpoints that can be filtered by query parameters. In particular filtering factoid-resources by statement content can lead to ambiguities: The result of the filter will return only factoids which contain statements that fulfill the filter criteria. Nevertheless, it would make queries easier, if the result could include information on the statements, sources and persons related to the factoids as well. This could be achieved by three methods:

> With a dynamic creation of the results, each object would include only the related resources that fulfil the filter criterias. However, this would entail an implementation burden and would result in "instable" factoids.
>
> The filter could be applied only to the queried resource but each of them could include all related resources even if not matched by filters. While this solution keeps the factoids stable, it also means that a query returns unexpected results.
>
> The resources in the return could only point to the related resources not giving any content. This slight modification of solution 2 ensures on the one hand that we do not return full content statements - even if the pointers are still present - that don't fit the filter criteria and on the other reduces the size of the result.

The factoid model is underspecified when it comes to cardinalities of the relationships. The typical situation probably is, that the researcher can extract several statements on a person from one source. But if the source talks about several persons, the factoid could make one statement on several persons (e.g. "X, Y, Z participated in the expedition to the South pole.").

The statement is the richest entity in the model as it should allow covering as many cases as possible. IPIF tries to identify a basic set of statements typically made on a person (name, memberOf, relatesToPerson, role, date, place). IPIF initially proposed a generic fall back property statementContent to cover other possible statements. This proved to be not specific enough, so IPIF introduced the distinction between statementType and statementText.

Strict conceptual models like the CIDOC CRM or BFO are based on a fundamental distinction between temporal and non temporal entities. This event based modelling seems to fit very well to prosopographical data, so other vocabularies like BIO (Davis/Galbraith 2010) enforces an event related modelling. Other models, like the BIO-CRM (Tuominen et al. 2018) or the TEI personography are less strict. Practice demonstrates that data often is not explicitly event based, e.g. with properties like "name".

A major use case for IPIF is autocompletion from external controlled vocabularies. This needs human readable identifiers for persons, like the typical controlled vocabulary entry as a combination of "name, claim to fame, birth, death". This solution does not cover the rich variance in possible descriptions in prosopography (Fokkens/Braake 2018): there are identifiable but anonymous persons (the bishop in a defined diocesis at a specific time, the father/wife of a person, etc.). "Claim to fame" depends on the context of data creation, birth and death are often not known, but a period of creative productivity is (floruit). The variety is well demonstrated by the results of a data-for-history-meeting in 2019 (Light et al. 2019). IPIF is about to propose a pragmatic solution.

The paper will present details on these and further modelling decisions, demonstrating how they affect the efficient use of the API and still can be reused in more strict conceptualisations of prosopographical data.

Bibliography

Althoff, Gerd. Möglichkeiten und Grenzen Elektronischer Datenverarbeitung bei der Erforschung der Geschichte des Mittelalters". Computers and the Humanities 12, Nr. 1/2

(1978): 97–107.

Barman, J., R. Barman, und W. T. Kirshaw. Prosopography by computer: the development of a database". Historical Methods Newsletter 10 (1977): 102–8.

Bradley, John, und Harold Short. Texts into Databases: The Evolving Field of New-Style Prosopography". Literary and Linguistic Computing 20, Nr. Suppl (1. Januar 2005): 3–24. https://doi.org/10.1093/llc/fqi022.

Bühlmann, Felix, Thomas David, and André Mach, "The Swiss Business Elite (1980–2000): How the Changing Composition of the Elite Explains the Decline of the Swiss Company Network", Economy and Society, 41.2 (2012), 199–226. https://doi.org/10.1080/03085147.2011.602542

Davis, Ian, und David Galbraith. BIO: A vocabulary for biographical information", 2010 2003. http://vocab.org/bio/.

Fokkens, Antske, und Serge ter Braake. Connecting People Across Borders: A Repository for Biographical Data Models". In BD-2017 Biographical Data in a Digital World 2017, 2119:83–92. CEUR Workshop Proceedings. Budapest, 2018.

Jackson, Cornell. Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland". Digital Scholarship in the Humanities 32, Nr. 2 (1. Juni 2017): 336–43. https://doi.org/10.1093/llc/fqv070.

Light, Richard, Torsten Hiltmann, Christopher Pollin, and Georg Vogeler. Working Group – Prosopographic Data, DfH 2019". Data for History - Forum, 11. April 2019. http://forum.dataforhistory.org/node/

Schwartz, Daniel. Syriac Persons, Events, and Relations: A Linked Open Factoid-based Prosopography". Utrecht, 2019. https://dev.clariah.nl/files/dh2019/boa/0875.html.

Tuominen, Jouni; Hyvönen, Eero; Leskinen, Petri: "Bio CRM. A Data Model for Representing Biographical Data for Prosopographical Research." In: BD-2017. Biographical Data in a Digital World 2017, ed. by Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, Eveline Wandl-Vogt, Budapest: CEUR (CEUR Workshop Proceedings 2119), 2018: 59-66.

Vogeler, Georg; Schlögl, Matthias; Vasold, Gunter; Stoff, Sebastian (2020): Prosopographische Interoperabilität – Stand der Dinge. In: Schöch, Christof (Ed.): DHd2020: Spielräume. Digital Humanities zwischen Modellierung und Interpretation. 7. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum. Paderborn u.a. DHd. 2020. 348-350. (and the poster itself at https://online.uni-graz.at/kfu_online/wbFPCompsCallBacks.cbExecuteDownload?pDocStoreNr=

Vogeler, Georg, Matthias Schlögl und Gunter Vasold (forthcoming): Data exchange in practice: Towards a prosopographical API. In: Fokkens, Antske (Hg.): BD2019. Biographical Data in a Digital World 2019. Proceedings of the conference held in conjunction with Recent Advances in Natural Language Processing 2019, Varna 5-6 September 2019. Wien, Budapest. CEU. Preprint: https://hcommons.org/deposits/item/hc:29017/

Warren, Christopher N., Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, und Cosma Shalizi. Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks". Digital Humanities Quarterly 010, Nr. 3 (12. Juli 2016).

Keywords: API, Prosopography, modeling