Using Markup for Multivariate Analyses in the Prosopographical Study "Formation for the Public Sphere"

Monica Langerth Zetterman
(monica.langerth@ped.uu.se)
Digital Literature, Uppsala University

Introduction

This paper aims to illustrate how markup might be applied for multiple purposes in research.[1] Here, the TEI/XML encoding scheme[2] was used as a research tool when producing a collective biography in a sociological prosopographical study on prominent Swedish female pioneers around the turn of the century 1900.[3] In this collective biography the markup is used for the exploration of biographical information. Although the markup provided was done for a special reason, namely to extract specific data in order to apply multivariate analyse methods, such as correspondence analysis[4], it also provides means for presenting, filtering and indexing the material.

Background

The main purpose of the project *Formation for the public sphere. A Collective Biography of Stockholm Women 1880--1920* is to investigate the social strategies of the first generations of women entering the public sphere in Sweden. This period was of crucial significance for women engaged in philanthropy, reform pedagogy, modern health care, literature and music. These women's strategies, investments and careers differed from their male contemporaries and their contributions are not easily recognisable. In order to discern and interpret their contributions to the establishment of the modern welfare state institutions, a modern educational system and the modern cultural fields, methods from the French sociological tradition founded by Pierre Bourdieu have been used.

A central endeavour is to collect information on the women's social origin, social intercourse, their networks, educational trajectories and matrimonial status. Such information is here called 'assets' or 'capital'. In Bourdieu's sense certain types of capital are acknowledged within certain social groups but not by everyone (Bourdieu, 1992). Each field that is sufficiently autonomous has its own rules for inclusion, exclusion and rewards, and specific species of capital.[5] By analysing the distribution of certain types of capital among the pioneer women we try to map the structure, the hierarchies and the polarities of domains like female culture, education and philanthropy.

Since we favour the collection of data which is sociologically interpretable data it is important to collect information on names, dates and places, e.g. where and when and with whom she lived, where she worked, in order to trace the "meeting places" and networks. Hence a mandatory core set of data was, whenever possible, harvested to depict some of the most crucial assets:

- Social origin: father's and mother's occupation, education, positions. Number of brothers and sisters. Woman's and parents' place of birth and place of upbringing.
- Educational capital: kind of basic and further education. Sojourns abroad.
- Social capital: influential relatives, matrimonial status, number of children, housing, member of state commissions, foundations.
- Economic capital: wealth, earthly goods and relations to patrons.
- Political and religious capital: positions in political/religious organisations, standpoints in such matters.
- Specific symbolic capital: assets being valued either within certain fields or domains or within women's networks.

Of course many of the biographical texts cover much more, but the main aim is trying to make the collection of these core data as comprehensive as possible for each woman.

Modelling the Data

There are two kinds of datasets of biographical accounts called capital descriptions. One of the sets consists of one hundred rather extensive texts written in running prose text by the researchers, aimed to be published in for example historical journals or biographical handbooks. The scholars have explored archival material such as letters, diaries or estate reports, as well as printed newspapers, journals or books — and of course existing biographies and autobiographies. The other set is more than 1200 condensed texts based on excerpted information.[6] The excerpts have been transcribed from two volumes, one from 1914 and one from 1921 with biographic articles on prominent Swedish women. Both kinds of texts have been provided with markup according to the TEI guidelines with the additional TEI tag set "Names and Dates" to encode proper names, date periods and precise dates.[7]

Similar to the much more extensive and ambitious *Orlando* project[8] we produce texts and we apply descriptive model-driven

and interpretative markup. Unlike the *Orlando* project, though, we have not developed a DTD for this project.

In our content model each woman corresponds to a main division that contains subdivisions and further subdivisions. In principle each subdivision or sub-subdivision corresponds to one type of capital such as . The basic features marked in a subdivision are:

- a date(range),
- an event or an occurrence,
- name(s).

For each of these elements there is a reciprocal relationship to a biography and an arbitrary category respectively. This means that a division is only instantiated when there is a relationship or an occurrence of a certain kind. Obviously, some instances are mandatory: we need for example to record a name and a birth date in order to instantiate a record for a woman. We use the hierarchies to extract exactly the data we need for the analyses. Thus, we can extract, the content "Paris" out of the element `placeName` when and only when it occurs within `div type="travel"`. We might thus extract information on that the woman in question did travel to Paris, and we do not have to bother with all other Parises that might appear: books published in Paris, dresses made in Paris, father born in Paris. Obviously a consistent use of the content model providing for the accuracy of the markup become very important since so much depends on it.

Our approach to apply the content model and do the encoding in running prose is not unproblematic. We have faced, and are still facing, many rather difficult decisions on the ways to encode and how to denote the aspects we want to capture. Should chronology take precedence over events or the other way around? The consistent use of divisions corresponding to the content model is important, since it guarantees that the internal hierarchy of elements is no hindrance for finding the specific data.

The corresponding types of data (i.e. the core set) from the two different datasets have been merged into one dataset for further processing. We extracted the data needed for analyses into tables using XSL in order to import the tables into the statistical software SPSS. SPSS is used for converting string values to numerical values and for organising, combining, classifying and aggregating variables. After preparations in SPSS the datasets were imported into the SPAD software were the correspondence analyses are done.

Since the material is heterogeneous, it calls for some measures to guarantee the consistency of data. If we had chosen a database solution, such as the relational database used by the project on the prosopography of the Byzantine empire, it would be a matter of course to opt for a controlled vocabulary, as do Bradley and Short. Thus the categorisation is undertaken prior to data input into order to ensure the congruency of the recorded aspects:

> Of course, this kind of interpretation of a source — by assigning some aspect to fit into categories — is in fact very similar to an important element of most scholarly work: classification and categorisation are standard part of scholarly practice.
>
> (ibid. 9)

This is very true also in our project. The major decisions on the categorisation have been taken prior to the collection of information and prior to the writing of the biographies.

Closing

What might separate this project undertaking from other similar prosopograhic projects is the aim to maximize portability. The material is available through an ordinary web browser, downloadable and reusable with intact markup. Users should be able to import the material to their own systems or add additional markup and not be forced to stay on our website in order to explore the material or perform their own analyses. Another difference is that the encoded data is used as input to multivariate quantitative analyses, as will be illustrated at the presentation. Thanks to the encoding we can provide enhanced navigation, presentation of different views of the material and filtering possibilities, which will be demonstrated in the paper presentation. Altogether we have found that the TEI encoding scheme has been serving as quite a valuable tool in this kind of collaborative research practice.

---

1. Parts of the content in this paper is based on work in progress, a forthcoming article co-authored with Prof. Donald Broady, titled "TEI markup as research tool in the prosopographic study Formation for the public sphere". In our collaborative work Prof. Broady answer for the sociological and historical content and the author of this paper for the markup, application and statistical analyses.
2. See `<http://www.tei-c.org/>` and Sperberg-McQueen & Burnard
3. See `<http://www.skeptron.ilu.uu.se/broady/sec/ffo.htm>`. The project is directed by Donald Broady and funded by the Bank of Sweden Tercentenary Foundation.
4. *l'Analyse des Données*, introduced by Jean-Paul Benzecri, a geometer-statistician, in the 1960. The method is done by modelling data sets as clouds of points in multidimensional Euclidian spaces and then interpreting the data in the cloud of points (Lebart et al.). Cf. Bourdieu (1984) for applications and some explanations.
5. See Broady for a proposed definition on Bourdieuan prosopography. See also the study on the French academic field *Homo Academicus* Bourdieu 1984) for an example of Bourdieu's prosopography.

6. Provided that the copyright issues may be solved, there should in due time be a freely available digital version. Meanwhile the access is restricted to the researchers and for teaching purposes.
7. cf. Sperberg-McQueen and Burnard, 2002, pp. 499-516 <http://www.tei-c.org/P4X/ND.html>
8. For information on the *Orlando* project, documenting "the scholarly history of women's writing in the British Isles." see <http://www.ualberta.ca/ORLANDO/>. See also e.g. Grundy et al.

Bibliography

Bourdieu, P. *Homo academicus.* Paris: Minuit, 1984. English translation: *Homo Academicus.* Polity Press, Cambridge, 1988.

Bourdieu, P. *Les règles de l'art. Genèse et structure du champ littéraire.* Paris: Seuil, 1992. English translation: *The Rules of Art. Genesis and Structure of the Literary Field*. Polity Press, Cambridge, 1996.

Bradley, J., and H. Short. "Using Formal Structures to Create Complex Relationships: The Prosopography of the Byzantine Empire--A Case Study." Ed. K.S.B. Keats-Rohan. Oxford: Unit for Prosopographical Research, Linacre Collage, 2002. Preprint available at <http://pigeon.cch.kcl.ac.uk/docs/papers/pbe-leeds>.

Broady, D. "French prosopography. Definition and suggested readings." *Poetics* 30 (2002): 381-385.

Grundy, I., P. Clements, S. Brown, T. Butler, R. Cameron, G. Coulombe, S. Fisher, and J. Wood. "Date ChronStructs: Dynamic Chronology in the Orlando Project." *Literary and Linguistic Computing* 15:3 (2000): 265-89.

Lebart, L., A. Salem, and L. Berry. *Exploring Textual Data.* Dordrecht: Kluwer Academic, 1998.

Sperberg-McQueen, C.M., and L. Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange (TEI P4).* Oxford, Providence, Charlottenville, Bergen: TEI Consortium, 2002.