The Biographical Portal of the Netherlands

Experiences in collecting and curating data from the web and elsewhere

Jelle GERBRANDY\textsuperscript{a,1} and Els KLOEK\textsuperscript{b}

\textsuperscript{a}Gerbrandy SRL, Torino, Italy, jelle@gerbrandy.com

\textsuperscript{b}Huygens Institute for Netherlands History, Den Haag, Netherlands

Abstract. The Biographical Portal aims to collect, organize, and make available all biographical information that is available about Dutch people. In this paper, we describe the design process, the choices we made, the tools we used that may be useful for a larger audience, and “lessons learned”: things that we feel we have done well, and things we would have done differently if we had known from the start what we do now.

Keywords. Biography, Standards, Cultural Heritage, Data Aggregation

1. Aim and Scope of the Biographical Portal of the Netherlands

The Biographical Portal (www.biografischportaal.nl) aims to collect, organize and disclose biographical information about people from the Netherlands. At the moment, it contains information about more than 80,000 persons, and links to more than 125,000 biographies from 26 different sources. It receives about 1,000 unique visitors a day, which is a number that is slowly growing.

The Biographical Portal does not only collect links to existing biographical information on the web but also:

- Digitize and publish biographical dictionaries from the 18\textsuperscript{th}, 19\textsuperscript{th} and 20\textsuperscript{th} centuries that were previously only available in printed form.
- Add and edit data about persons. For each person in the portal, we aim to offer at least structural information about date and place of birth and death, name variants, field of interest and some basic categories (religion) available.
- Identify different biographies of the same person.
- Offer search tools to make the information discoverable.
- Assign a unique number, and offer a permalink, for each person.

The Biographical Portal aims to become the reference point for biographical information about Dutch people.

The Biographical Portal is the result of collaboration between 10 research institutes that are active in the humanities and the cultural heritage sector. The Huygens Institute for Dutch History is the host for both the website as for the daily management of the portal, and has committed itself to the long-term hosting of the project.

\textsuperscript{1}Corresponding Author.
2. Methods Used

2.1. Collecting existing information from the web

A first inventory found about 40 different collections of biographical information on the Internet that are of sufficient size and quality to merit inclusion in the biographical portal. At the moment of writing, the biographical portal contains information about the biographies in the 20 largest collections on-line collections (the other sources for the biographical portal are digitized books).

Some of these collections are owned by one of the 10 institutes that are involved in the project, others are from third parties. From these data providers, we ask not access to their data, but also an investment (in the form of time) for arranging the conversion of data. Sometimes, that needs convincing. We point out that also they gain from inclusion in the biographical portal: availability in the biographical portal improves the visibility and discoverability of their original data.

A second reason to cooperate is transfer of knowledge. By converting its data into a standard format, an organization can be helped to improve their practices for representing their own data. At the same time, the biographical portal offers its partners an opportunity to improve the quality of their the contents themselves – the information that we add in the biographical portal by identifying and editing the data (see below) can flow back to our partners.

It is important to note in this respect the Biographical Portal does not publish biographical texts itself – it functions as an index and a search engine rather than as a publisher of actual biographical texts. This means that we are not in competition with potential providers of data, but offer a complementary service.

2.2. Digitization

Digitization of biographical dictionaries provides a specific challenge compared to “ordinary” books – the dictionaries contain highly structured information, organized in entries per person, perhaps with a portrait to each entry and essential (genealogical) data, which isn’t as easy to digitize as for instance an ordinary textbook.

Digitization of existing biographical information was done by using existing software developed by the Huygens Institute of Netherlands History. Details fall outside the scope of this paper; more information about this project can be found at www.historici.nl/pagebrowser.

2.3. Ingesting of data

Importing of data from our sources works as follows. The data is first converted into a data exchange format in XML. For this purpose, we defined a TEI-variant called “BioDes” (www.biografischportaal.nl/about/biodes) which is described in more detail below. Depending on the technical knowledge and availability, some of our partners did the conversion in-house; for others, we did the data conversion ourselves.

Once ready and validated, the BioDes data files are imported in the biographical portal. This import is idempotent and can therefore be re-run when we want to. This is important in the case of data sets that are still in development: some of our data providers are constantly editing and adding their biographies, and it is important have a mechanism that allows for re-importing data that allows for this possibility.
Imported data is immediately published: there is no intermediate editing stage before new data is rendered public. We do curate and edit our data (more about that below), but that is done while the data is published already, and, in a sense, “publicly”. We are happy with this choice of policy, because it reduces drastically the time and the number of people involved when ingesting new data. A drawback, though, of “publishing before curating” is that the data that we publish is subject to constant change. For example, a fact that is often hard to explain is that during the editing process, when biographies are marked as being of the same person, the number of “persons” in the biographical portal may actually decrease over time.

2.4. BioDes

The quality and structure of the data from our sources vary considerably. For some sources, the available data consists of little more than a name and a link. On the other end of the spectrum are sources that include images, detailed information about life events and relationships to other persons.

We tried to design our data representation format, BioDes, to be as flexible as possible, so it can represent both very rich as well as very minimal data sources.

For inclusion in the Biographical Portal, a BioDes document needs to contain just a few elements: the name of the person, a URL of the place where the biographical information can be found, and the name of the publisher. This minimal case applies, for example, to digitized biographical dictionaries, where just the index of the dictionaries (e.g. a list of names and page numbers) was available.

On the other end of the spectrum are very rich sources and annotated resources. For example, the Parlementair Documentatie Centrum makes biographical information about politicians available. This is a data set that contains a wealth of highly structured information, and contains, for example, data about memberships of parties and significant events. The TEI tags event and state proved to be useful general-purpose tools for structuring these types of information.

Another property of the kind of data we are dealing with – historical data about persons – is that there is often a certain amount of insecurity in the data: the exact date of an event may not be known, or be known only approximately. The flexibility of the TEI tag set has been very useful in this respect. When we defined the BioDes format, we only had a general idea about the structure of the data sets that we wanted to include, and could not predict all the different use cases that would occur. In particular, lack of knowledge about precise life dates can, in historical biographical data, take a large number of different forms - “born in the second decade of the 15th century, but surely before the 15th of may 1418” is a case that we did not foresee, but that TEI handles gracefully.

BioDes is not only used as an intermediary format for importing data, but is used throughout the whole application to represent our data. When our editors add new information, they are (behind the curtains, of course) effectively creating BioDes documents.

2.5. Identification and editing

After importing a new database, the following two steps are taken:
1. We add basic biographical data such as dates of birth and death, if this is not available yet in the original data source.

2. We identify biographies of the same person.

This is a laborious task, and our editors are being assisted as much as possible by software tools.

The identification process is partly automatized. We created an algorithm that assigns a “similarity score” to any two biographies in the portal. This score weighs a number of different features – similarity of the names, time of birth and death, etc. Especially the comparison of names is complicated by the disparity of the data sources. Surnames are not always marked as such; sometimes only initials of given names are present, etc. For example, the algorithm needs to recognize that “C. Huygens”, “Constantijn Huygens” and “Huygens, Constantijn” are all quite similar, but that “H. Constantijn” is almost surely a different person.

Those pairs of biographies that have a very high score are identified automatically. Pairs of biographies with a lower score are checked by hand. For this, our editors are presented with a list of pairs of biographies with their assigned score, and can quickly select those pairs that are actually of the same person.

This procedure is relatively quick, and seems to work well, although without a reference corpus or “gold standard” it is difficult to measure the effectiveness of the identification process in a precise way.

The tools we use for editing are more traditional – our editors work with standard web forms that have been designed and improved in close collaboration between the editors and the programming team. A very useful feature in the editing process is the automatic detection of discrepancies between our data sets – it often happens that two biographies will have conflicting information about, say, place of birth. This often point towards an error, and is a signal for our editors to check, and if possible improve, the situation.

2.6. Combining data sources

When two or more biographies are marked as being about the same person, a new question is raised: which information do we show on the web site as being the information about this person?

The biographies may have partial information (for example, no exact date of birth) and complementary information (one biography list pseudonyms of an author, say, while another mentions her maiden name). Also, two pieces of information may plainly contradict each other.

Our approach is this. We have ordered our data sources by “trustworthiness.” Data that is added by our editors is highest on that list. For each piece of information (say, a birth date), we take the most data of the most trustworthy source we can find. This is a fairly straightforward and transparent way of combining our data sources; it presents our end users with data with a merged result that is as complete as possible, allows us to resolve contradictions automatically, while at the same time it gives our editors complete control over the data that is being shown.
2.7. Search

The biographical portal contains information from a large number of diverse sources, and this information needs to be organized in a way that our users can find the information they need.

Two aspects of the search problem are particularly relevant when dealing with data about people: searching for names, and searching for dates.

Names, especially in a historical context, present specific problems with regard to orthography: the same name can and will be spelled in many different ways. A standard solution to the problem is to simply list all orthographic variants of the name. This approach as much to say for it, but is also very labor intensive, and is simply not possible to do well with our data set and within our budget. Instead, we have tried to create an algorithm that tries to match names on the basis of phonetic criteria. (The same algorithm is being used in the similarity algorithm described above).

This algorithm for fuzzy search of names has been custom-made for the biographical portal, but can be (and in fact has been) re-used in other projects as well. The idea is based on that of SOUNDEX: each word is reduced to a (or possibly several) simplified “phonetic representation” that represents, more or less, what the word sounds like. These phonetic representations are then based as the basis for searching and comparing.

For our use case, directly using the soundex algorithm was not an option: soundex is based on English, and does not give very good results for the Dutch case, which, as any other language, has its own peculiarities. In any case, soundex would be too rough a measure for our data set – using soundex would give too many false positives. Instead, we developed an algorithm ourselves that tries to find a balance between finding all relevant results while not finding too many irrelevant ones. The algorithm is still being tweaked and improved upon on the basis of user feedback.

Another search problem concerns the temporal aspects of personal data. Say, a user searches for all people that lived in the 17th century. Clearly, the results should include anyone that was born or died in that time. But in many cases, the exact date of birth is not known (but specified as a period), or only a date of baptism or funeral is available. For example, a person with an unknown birth date that we know to have been buried between 1715 and 1720 will probably have been alive before 1700, and should therefore be included in the search results.

3. Lessons Learned

3.1. Copyright issues

From the beginning, the idea was to make our information available freely, probably under some Creative Commons variant. However, this idea was never formalized, and no written agreements with our data providers have been made. We confronted the problem of defining a license for our information when the portal was already operative and the data from our sources already published. So now we are in the unfortunate situation that we need to negotiate a compromise with our partners about the license. This was clearly a problem that we could have avoided if we had confronted the issue in an earlier stage. The question of copyright is further complicated by the fact that the
biographical portal does not simply reproduce content of other parties, but combines and changes that information as well.

3.2. Standards

The BioDes XML definition plays a central role in the data management of the biographical portal. Although we are very happy with the internal functioning of these data structures, there is one thing that we would have done differently. While most of the specifications and actual tags in BioDes format are taken from TEI, it is not the case that BioDes documents are valid TEI documents. The reason for not following TEI to the letter was one of simplification (TEI is very much geared towards document meta-data, while we were interested in data about persons). But deviating from the TEI standard comes at a price: not only the obvious one of the fact that our data representation is not natively compatible with other services that use TEI (which is a technical issue that can be resolved relatively easily), but also because strict TEI-conformance is a strong argument for other organizations to adopt our standards.

3.3. The trade off between quantity and quality

On the one hand, we want the portal to be as complete as possible – we would like to include any data that is of sufficient quality. More is better. Now that the procedures and software are in place, ingesting new data sets is relatively fast and cheap. Also, the inclusion of a new dataset is always an easy “photo-op” - it represents a clean milestone that can be presented to the management as a success, or can serve as a topic for a press release.

On the other hand, we want to data to be usable and as trustworthy as possible. For that reason, complete and correct meta-data is essential. However, the process of identifying persons and adding and correcting basic data about individual persons is slow and costly. Moreover, this part of the process is more difficult to “sell” - small incremental steps in the improvement of the data themselves are all but invisible to the average user and, in the light of the relatively high costs, it is more difficult to convince management that these funds are well-spent.

The decision of adding new content to the repository always requires a trade-off between these two aspects. Up until now, the policy of the biographical portal has always been to try to not let the quality of the data suffer by including un-curated new data sets – in other words, new data are only added if time and money for editing these new data is available, or if they come with complete and high-quality meta-data themselves.

3.4. Indexing versus content

This is an issue that is specifically relevant for the project at hand. The decision not to include the biographical texts themselves in the portal rendered it very easy to make agreements with our data providers, who in this way only benefited from inclusion in the portal. However, the lack of original content in the portal also works against us: it makes the portal less visible on the web than it could be. We provide only essential data for each person, with links to places where further information can be found. This implies that for search engines like Google, the pages we refer to will often rank higher than our pages – as those are the places where more detailed information can be found.
A similar issue holds for human editors, such as those from Wikipedia, that prefer to refer to original sources rather than to a pages on a portal such as ours.

4. The Future

At the moment, many of the most important collections of Dutch biographical data have been added to the portal. We aim to continue what we are doing by adding and integrating new collections of data. But we have other plans, for which we are seeking partnerships and funding:

- Creating new biographical content (that is: writing new biographies)
- Adding more detailed data about persons, such as relations between people, more detailed information about life events, information about membership of groups, etc. Some of our data providers already give us this information, but at the moment we have no way yet of presenting this information to our users. Adding information about relation between people is already difficult do to well within a single dataset, but represents a completely new challenge when data-sets need to be merged that use different concepts and different ways classifying relationships or events.
- Creating tools to help (historical) researchers find and organize the data. A special challenge here is in creating tools that are useful – tools that help researchers with specific research questions, but that can be generalized to be of use to other researchers as well. Close collaboration between researchers and programmers is a precondition for succeeding at this, and finding people from both camps that are willing and able is not easy.
- Making our information more visible on the web, for example by creating ways of making the information more attractive for a larger audience. For example, we are planning a national portrait gallery with extensive biographical information about 2,000 important Dutch historical characters. Another ambition is to connect our data to a larger international context (a “European Biographical Portal”). Editing data by hand is very labor intensive, and technology can, at this stage, be of only limited help. We would like to create a community of volunteers to help with editing data.

5. References

- The Biographical Portal of the Netherlands: [http://www.biografischportaal.nl](http://www.biografischportaal.nl)
- The Huygens Institute for Dutch History: [http://www.huygens.knaw.nl/](http://www.huygens.knaw.nl/)
- BioDes XML standard: [http://www.biografischportaal.nl/about/biodes](http://www.biografischportaal.nl/about/biodes)