Biography Ned

Funded by the eScience Center
http://esciencecenter.nl/

Serge ter Braake (Fac. of Arts, History)
Antske Fokkens (Fac. of Arts, Comp. Linguistics)
Niels Ockeloen (Fac. of Computer Science)

VU University Amsterdam and Huygens ING

Project leaders: Guus Schreiber (Faculty of Sciences), Piek Vossen and Susan Legêne (Faculty of Arts), together with Els Kloek (Huygens/ING).
BiographyNed: Extracting relations between people and events

The Biography Portal of the Netherlands links more than 75,000 Dutch people mentioned in various databases, through a limited set of metadata. This project aims to enhance its potential for historical research by transforming the available data into a semantic knowledge base and through the creation of a demonstrator. BiographyNed is a multidisciplinary project that combines expertise from history, computer science and computational linguistics.

The lead questions for the design of a semantic demonstrator are: Which relations can we reveal between people and events, geographical movements and networks between people? What do they tell historians about the formation of Dutch society and the ‘boundaries of the Netherlands’?

The current search engine lacks the analytic tools to show interconnections, trends, geographical maps, time lines, etc. This project aims to strengthen the value of the portal and comparable biographical datasets for historical research, by improving the search options and the presentation of its outcomes, starting from the Simple Event Model. The demonstrator will add a semantic layer on to the current Biography Portal. This layer can also include information from external resources, such as museum objects or Wikipedia. Ultimately, the project may help to reveal unknown relations between people and events by linking data that has mainly been studied in isolation so far.

Guus Schreiber – Project Leader

Enhanced Science = Linking the world of history

The BiographyNed project aims to create a semantic knowledge base by extracting relations between people and historic events in the Biography Portal.
Goals

- eScience centre: supports and reinforces multidisciplinary and data-intensive research through creative and innovative use of ICT in all its manifestations.

- Biography Ned:
  - Develop generic *technologies* AND *methodologies* for e-Humanities
  - Reproducable, modular and extendable
  - Applied to the data of the “Biografisch portaal“ created by Huygens/ING
  - From the perspective of historians as researchers
This is the site of the Dutch Biography Portal. You can find scientific information here about various leaders and other prominent figures from Dutch history, from the earliest times up to the present day. The website will be constantly replenished with information and images.

News

- Due to maintenance works the Biography Portal is currently unavailable
- Recently added: the Compendium of office holders and civil servants (1428-1861)
- Recently added: the National Biographical Dictionary (NBW)
- Recently added: the Magistracy of Amsterdam (1578-1795) by Elias
- New search option: religion
- The data of the portal are downloadable using the BioPort API
- More than 110,000 biographies!
Various data sets linked through shared meta data

Search

- full name
- family name
- born
- place of birth
- died
- place of death
- alive
- gender
- category
- religion
- source
- free search in the text
- bioport identifier

Examples:
- e.g. huygens or aalb*.
- e.g. may 1650 to september 1655.
- e.g. Zaamslag.
- e.g. before 1914.
- e.g. Den Helder.
- e.g. 1795 to 1813.
- e.g. vroedvrouw or *moord*
Why a good case for eScience?

- Involves big data with high complexity
- Rich meta data joining diverse textual sources and selections of data
- Incomplete and noisy
- Potential to investigate difficult questions, e.g.:
  - How did the current Dutch elite develop from the colonial past
- Biographies may represent different views and realities and thus answers to questions:
  - hero or villain
  - 2.8 textual sources per person
Statistics on available information

Individuals with available information (%)
Availability of Information in the portal

![Bar chart showing the availability of information in the portal. The x-axis represents different categories such as Partner, Mother, Father, Claim 2 Name, Religion, Occupation, Date of birth, Place of birth, Date of death, Place of death, Category, and Name. The y-axis represents the number of records. The chart compares the 'Information Absent' and 'Text available' for each category.]
What will we do?

- Develop generic text mining technology that converts textual data to structured data:
- Develop rich, more completed and formally structured data repository
- Develop visualisations and interactions on the data set to support historical research
- Develop a range of cases that demonstrate the possibilities and impossibilities of the data set and technology
Challenges for text mining

- Different genre for NLP (more creative styles): e.g. unknown effects on detecting named-entities
- Old variants of Dutch
- Special types of data: education, careers, kinship, historical events
- Complex types of data: event perspective, event value and opinion, intentions, causations
- Contrast different narratives of the same person that only lived one life
Reproducibility/Replication

• What do results mean if they cannot be reproduced?
• What variation in results can be expected based on details not mentioned in papers?
• Which information is needed to replicate results or find the origin of differences?

Paper submitted ACL 2013 (joint work with Marieke van Erp and others)
Representation (events)

- How to combine knowledge from the NLP community and Linked Data community?
  - Combination of textual information with external resources
  - Complete representation of information from text (location, retrieval method)

Paper submitted to workshop on Events: Definition, detection, coreference and representation (joint work with Marieke van Erp, Willem van Hage, Sara Tonelli, and others)
Challenges for web technology

- Conversion of the data to formal semantic models.
- Identity of persons: who is who in the portal and the web?
- Linking to external sources (Linked Open Data)
- How to deal with provenance of data derived from sources?
- How to handle biases in selections (by sources and on available sources).
- How to design a demonstrator: balance between: easy, intuitive, appealing, precise, scientific, flexible, open and extendable, innovative.
Johan Rudolph Thorbecke werd in 1798 geboren op 14 januari in Zwolle en komt uit een half-Duitse...
The Art and Science of History: Drawing up a narrative from primary and secondary sources which approximates historical reality as well as possible.
Building Blocks and Concrete

- Building blocks: facts derived mainly from archival findings and existing literature
- Concrete: the methods historians use to put them together into a narrative/synthesis.
- The Narrative: a historical synthesis which cannot be scientifically proven (only made likely) based on facts which can be proven or falsified. There is necessarily a creative element in drawing up a narrative.
The House of History

A: Building Blocks

<table>
<thead>
<tr>
<th>Number of battles won</th>
<th>Hatred for dogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love for women</td>
<td>Public Opinion</td>
</tr>
<tr>
<td>Inferiority Complex</td>
<td>Death of friend</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>Career</td>
</tr>
<tr>
<td>Network</td>
<td>Lost at Leipzig</td>
</tr>
<tr>
<td>Lost in Russia</td>
<td>Place of Death</td>
</tr>
<tr>
<td>Parents</td>
<td>Date of Death</td>
</tr>
<tr>
<td>Wife</td>
<td>Children</td>
</tr>
<tr>
<td>Place of Birth</td>
<td>Date of Birth</td>
</tr>
</tbody>
</table>

B: Concrete

<table>
<thead>
<tr>
<th>Theory</th>
<th>Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic</td>
<td>Comparisons</td>
</tr>
<tr>
<td>Analyses</td>
<td>Reasoning</td>
</tr>
<tr>
<td>Interpretations</td>
<td>Quantifications</td>
</tr>
</tbody>
</table>

C: House

Synthesis

Conclusions
Our Sources Here

- The Metadata: building blocks
- The entries in biographical dictionaries themselves: short historical narratives
Where do eScience and History meet? (I)

A. Quantitative analyses of a larger group of people (prosopography). Surpassing the anecdotal.

B. Finding relations/networks between people which are otherwise hard to detect.
Where do eScience and History meet? II

C. Insight in *historiography* and historical *selectivity*. Who was described/included and why?

“Undoubtedly I have deprived many interesting women by not including them. The only thing I can say to defend myself is this: history writing is also a process of ruthless selection.” (Els Kloek, Head Biography portal and main author 1001 vrouwen)

D. Thematic research. E.g.:

When did the discovery of America start to influence people’s lives?
Case: The Dutch Nation & Identity

- Who were selected to be included in National Biographical Dictionaries and why? (what was their claim to fame?)
- Are there different perspectives on the same person over the time and how can this be explained?
- Who was deemed most important? (based on the length of the entries)
- What time periods are most represented?
- Is there a difference in claim to fame for people from different periods in history, or between men and women?
- Which words are used most often and can we link them to national identities?
Case: More Questions ...

- What events are mentioned most often and what does that say about the status questionis of how the Dutch see/saw themselves?
- What are the differences in the answers to these questions between several national biographical dictionaries?
- Are people and events described or appreciated differently over time? Does the perspective change?
- How does this relate to biographical dictionaries, nations and identities elsewhere in Europe?
Questions