# How to Make it in History. Working Towards a Methodology of Canon Research with Digital Methods

**Serge ter Braake and Antske Fokkens**

History and Computational Linguistics, VU University Amsterdam
De Boelelaan 1105 1081 HV Amsterdam, the Netherlands
s.ter.braake@vu.nl, antske.fokkens@vu.nl

**Abstract**

This paper proposes a methodology for studying canonisation of people in history with digital methods. These canons are for the most part culturally determined. For a select group of people, there is no doubt that they merit the necessary attention, but there is a large gray field of 'second rate' individuals who had an impact on history of which only a small group is included in more than a footnote. This makes the attention people get from historians rather arbitrary, subjective and unacademic. Digital humanities technologies can help us to work around this arbitrariness and to get insight into the canonisation processes.

**Keywords:** Canonisation, Fame, Ngrams, Named Entity Recognition

## 1 Introduction

This paper proposes a methodology for studying canonisation of people in history with digital methods.[1] With 'canonisation of people in history' we mean the repeated mentioning of people in any history book (e.g. a study on British Parliament), reference work (e.g. a biographical dictionary), newspaper, website or actual canon (e.g. the 'Canon van Nederland'.[2] Canons are for the most part culturally determined, rather than by the actual impact people had in history. The example of the continuous underrepresentation of women in history works makes this only too clear (Bosch, 2014). For a select group of people, there is no doubt that they merit the necessary attention in historiography, but there is a large gray field of 'second rate' individuals who had an impact on history of which only a small group is included in more than a footnote. This makes the focus of historians on a relatively limited group of people rather arbitrary, subjective and unacademic. Digital humanities technologies can help us to work around this arbitrariness and to get insight into the canonisation processes.

In this paper we take canonisation of individuals in the Netherlands as our example, but the same methodology could be applied to other countries. The rest of this paper is structured as follows. In Section 2, we introduce the phenomenon of canonisation in history and the role digital humanities can play. Section 3 discusses the different sources that could provide an answer to our question. In Section 4, we provide a breakdown of the available biographical data and tools for the Netherlands, how to make good use of them and what their limitations are. We propose a methodology for making the best use of digital methods in combination with traditional methods for canon breaking research in Section 5. In Section 6 we show some preliminary results, which is followed by our conclusions.

[1] All URLs in this paper were latest retrieved on 31 May 2015
[2] `http://www.entoen.nu/`

## 2 Canonisation and Digital Humanities

Canonisation of people and events in history is an unfortunate, but natural process. Once individuals are mentioned and remembered in various sources, they enter the frameworks people use to maintain their memory for a longer period of time (Halbwachs, 1985, p.29). This means that once well-embedded in collective memory or historiography, a person does not leave it easily and that those that did not make it are doomed to oblivion, unless they are (re)discovered. The urge to make formalised 'canons' of what everyone should know about history, no matter how useful for education and public history, reinforces this process. This means that historians could be 'blind' to large groups of potentially historically interesting people and events. Canonisation therefore impedes historical innovation and it needs to be studied in order to break it.

The problem of biases in historiography are well known, but there has been little research into how selection processes work and what this could mean for our knowledge and views of history as a whole. For the historian, this effect of reinforcing what we think to know about history and continuously forgetting/ignoring what we do not know poses a major, and as yet still underestimated, problem (Sample, 2012; Earhart, 2012).

One of the main challenges in addressing this problem is that identifying influential people that did not make it into the history books is a process of collecting needles from a haystack. Historians need to go through vast amounts of data that contain references to influential people and find those people that are forgotten despite being equally influential as their famous or semi-famous contemporaries. Digital methods are necessary to carry out such research in an efficient way.

The advent of the digital age has in general sparked a new interest in frequency lists, which help us in understanding canonisation processes. Ngram viewers can tell us the frequency of a (combination) of words within a certain corpus of texts over time, we can count the number of words used by members of parliament and the kind of terms they use

and we can evoke fame rankings of people who are mentioned in Wikipedia.[3] Such lists are particularly interesting for humanities researchers, since they give them the opportunity to approach old topics in a different way.[4] Computer software is able to analyse much more text than any human could ever do, which allows humanities researchers to back up interpretations based on anecdotal evidence with actual numbers and to formulate or test hypotheses more quickly. With the Google Ngram viewer, based on the words in millions of books, it is for example easy to see how the popularity of Anne Frank rises quickly after the Second World War.[5]

The creators of the Google Ngram viewer have run some interesting experiments with their corpus (Michel et al., 2011). The most closely related to our goals are the ones on the rise to fame of all famous people between 1800 and 2000 and the 'Science Hall of Fame.'[6] The first experiment used the 740,000 names of persons in Wikipedia and 42,358 names in the database of the Encyclopedia Britannica. This yielded interesting results, e.g. 1) Most people knew a quick rise to fame followed by a slow decline after the peak; 2) Most people enjoyed their peak circa 75 years after their births; 3) People increasingly become more famous more quickly, but also are forgotten more easily (Michel et al., 2011, p.180).

Online biographical dictionaries and Ngram viewers give ample possibilities for investigating who became famous and why, even when taking all the source biases and limitations of the tools into account. It is more challenging however, to look for the people who did *not* become famous, while they were prominent enough in their own time. Even if the data in Google Books and the KB Ngram viewer are less discriminative than the biographical dictionaries, they do not solve this problem. When the creators of the Google Ngram viewer did their research on the fame of people between 1800 and 2000 they used *existing* lists of people from Wikipedia and the Encyclopedia Britannica. Even if the lists from the Encyclopedia 'reflect a process of expert curation that began in 1768' (Michel et al., 2011, p.180), it still is biased and subjective. Logically, the people who are left out of Wikipedia and the Encyclopedia do not show up in the fore mentioned two analyses either and therefore, to a certain extent, the canon reaffirms itself.

These experiments are, in other words, top-down: existing lists were used to match with records of the past. The experiment can show that certain people are not mentioned as much as one would expect, but not that certain people or events were 'hot topics' during a certain time, but have been forgotten since. For a complete picture, records need to be queried that do not have the bias of modern records. We want to scan for any names in a wide variety of not only books, but also sources like journals, newspapers, pamphlets and archive material, and see what happens to their fame in the course of centuries. In Section 3, we will say a bit more about the potentially interesting sources to use to get a grasp on these 'missing persons.'

## 3 The sources

To map canonisation in history we need to make a distinction between the different sources we can use. There are contemporary sources (e.g. a pamphlet from 1581 scolding William of Orange) on the one hand and sources written after the death of a person (e.g. a biography on William of Orange from 1978) on the other. Similarly, there are sources with a conscious selection of people (e.g. historical sources like a biographical dictionary) and sources that do not or less consciously select (like a list of land owners). Obviously we can have both contemporary and later sources with and without a conscious selection, as can be seen in Figure 1.

The contemporary sources are needed to see how famous a person was in his or her own time. We will see in Section 6, Table 2 for example, that the politician Johan Rudolph Thorbecke was extremely prominent in the newspapers of his time. It is logical to assume that a person is often most famous in his or her own time, but the examples of the painter Vincent van Gogh and Anne Frank already show that this is not always the case. The sources published after the death of an individual show how the fame of a person developed. Even though Thorbecke remained one of the canonised figures from Dutch history, his fame declined over the years, as can be seen from the sources after his death. Obviously, for historical figures before the nineteenth century this starting point will be difficult to determine, due to the lack of sources.

It is more complex to make a distinction between sources that consciously select people to write about and sources that do not. A biographical dictionary is a good example of a source that *does* consciously select individuals. One of the main questions of any editor of a biographical dictionary is who is noteworthy enough to get an entry and who is not. A history book on the Dutch Revolt is already a less clear example of selection. Obviously, any historian selects the people and events he or she deems important enough to describe. The mentioning of individuals might, however, have to do with the selection of an event (e.g. presence at a certain battle) rather than with any selection of persons. This is why we consider prosopographical studies, group biographies, as good examples of sources that do not consciously select individuals. Prosopographies are quantitative studies on larger groups of people. The *category* a person belongs to (e.g. officers at the Council of Holland) determines whether someone is selected for the study, not the person him- or herself. Newspapers also select what they deem the most important news, but that is mostly driven by popular demand and not by historical judgments on who is influential enough to include.

The differences between these sources have to be taken into account for any historical interpretation of the results. In

[3] Wikirank: `http://wikirank.di.unimi.it/index.html`; Pantheon: `http://pantheon.media.mit.edu/methods`

[4] e.g. When was the word potato used for the first time? 'De DBNL ngram-viewer van de KB': `https://www.youtube.com/watch?v=XpMqypF46RY`

[5] `https://books.google.com/ngrams/` search for 'Anne Frank', on 13 May 2015.

[6] `http://www.sciencemag.org/site/feature/misc/webfeat/gonzoscientist/episode14/index.xhtml`

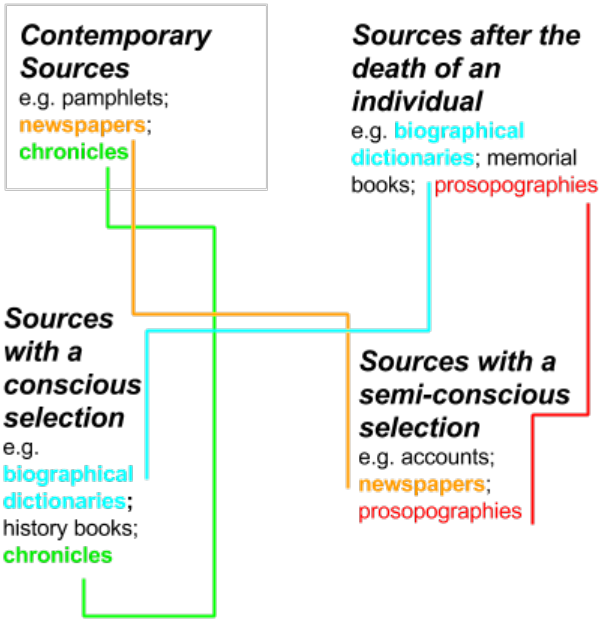


Figure 1: Schematic view of sources for research on canonisation processes

the following section we shall see how this already is facilitated.

## 4 Available tools and data

To investigate who became a well-known person, who did not, and why, we need at least the following data from as many records as possible: names, dates and places of birth and 'claims to fame' (i.e. why did or could someone become famous?). In theory any record of people or events could be suitable for our purpose, from medieval chronicles to early modern newspapers, to modern school books. For a full picture of canonisation a wide variety of sources needs to be consulted, from each category listed in Figure 1. Lists of famous people are strongly dependent on the kind of medium that is consulted, as will be demonstrated in Section 6. The enormous amount of data from these sources could never be close-read by one person. We therefore need digital methods to speed up our research (Wilkens, 2012, p. 251, 255), (Michel et al., 2011, p. 176). In this section we will discuss a non-exhaustive selection of what we deem to be the most obvious sources to start such research, and how they relate to the sources mentioned in Section 3.

It is relatively easy to trace the people who *did* make it in history for a top down analysis of canonisation. Biographical dictionaries list the supposedly most noteworthy men and women from, for example, a country, profession, time period or political movement. Many countries host a dictionary of national biography online, offering increasingly enhanced options for research.[7] People described in biographical dictionaries were selected by the editors, often after large consulting rounds. Sometimes the availability of experts also had an influence on who is included and who not (Hanssen, 1995, p.78), (Nadel, 1984, p. 52). In the Biography Portal of the Netherlands, 23 of such biographical datasets are gathered,[8] resulting in biographical data on over 75,000 individuals. These individuals can be analysed on common characteristics, such as age, gender and claim to fame. The dataset of the BP is used in our bottom-up analysis in Section 6. These biographical dictionaries are excellent examples of sources with a conscious selection after the deaths of individuals. Because, if all is well, individuals only have one entry in a biographical dictionary, their fame can be 'measured' by looking at the occurrence in other people's biographies.

Resources such as DBpedia,[9] a structured dataset in RDF based on the data in Wikipedia, offer similar possibilities for group analyses of 'famous' people. The advantage of these datasets over biographical dictionaries is that they are bigger, dynamic, more inclusive and edited by 'the crowd' rather than by a selected group of editors. Wikipedia does particularly well in providing reliable basic data on individuals. One of the disadvantages is that DBpedia and Wikipedia have clear biases as well, which are more often grounded in 'Geek hobbies' than in academic tradition (Rosenzweig, 2011). The broad criteria Wikipedia uses for inclusion nevertheless make it a source with a less conscious selection. Furthermore, it provides continuously updated information on people, both during their life and after their death (e.g.: actor Leonard Nimoy (†2015) had an extensive entry on Wikipedia during his life, which is still being adjusted and complemented as we speak.[10]

For data on Dutch people that were even less consciously selected the KB (National Dutch Library) Ngram viewer is a good resource to start.[11] The main advantage of the KB Ngram viewer is that it uses the words in over 9 million digitised newspaper pages from the Netherlands and thereby also covers people and events that were once considered worth mentioning and might have been forgotten in historiography. Unfortunately, the biases which are introduced by the limited availability of digitised newspapers will also influence the results provided here.[12]

The data derived from Google Books and made accessible in the Google Ngram viewer and its raw datasets are less specific for the Dutch situation, but still useful. They are less biased by preselections of digitisation than the newspaper archive. The 'black box' of the Ngram viewer, however, makes it impossible to see to what extend sources with a

[7] e.g. Oxford Dictionary of National Biography: http://www.oxforddnb.com/; Deutsche Biographie: http://www.deutsche-biographie.de; Australian Dictionary of Biography: http://adb.anu.edu.au/

[8] http://www.biografischportaal.nl

[9] http://www.dbpedia.com

[10] https://en.wikipedia.org/?title=Leonard_Nimoy

[11] http://kbkranten.politicalmashup.nl/

[12] This point was also made clearly by Bram Mellink in his presentation at 'studiedag God in Nederland 3.0' (21 november 2014) entitled 'Zoekt en gij zult vinden. Digitale onderzoeksmethoden, religiegeschiedenis en het probleem van de ondoorzichtige dorpsrel (1951-1952)' Slides: http://www.religiegeschiedenis.nl/rg/docs/PresentatieBramMellink.pdf

conscious or less conscious selection of people were used. The Ngram viewer calculates the word frequency in a selection of 5 million out of the 15 million books scanned by Google. The Ngrams are available in corpora of several languages, though not in Dutch (Michel et al., 2011).[13] Furthermore, there are Google NGrams for Dutch, which is a dataset of 133 billion words extracted from open websites between October and December 2008[14] and the DBNL Ngram viewer, which searches in Dutch literary texts.[15] For this paper we have used all these four Ngram viewers.

## 5 A Methodology for Canon Research

In this section, we propose a method for fruitful computational analysis of canon formation with digital historical data. As mentioned above, Ngram viewers are suitable for 'top-down' research on canons, when you know which people you are looking for. We want to combine this approach with a bottom-up approach, where the starting point is not an existing list of names, but *all* the names from as many resources as possible from all categories as described in Section 3. This way we can also find whose fame did *not* last for centuries and formulate ideas on why this is the case.
Another identified problem with Ngram viewers is that they provide little context and provenance information. Especially for a historian, it is important to know where information came from, to check the reliability and to see the context (Fokkens et al., 2014). We therefore need to facilitate the need for provenance and context by making a division between the original data and a layer above the original data (a supraset) where computational reasoning has taken place. Both the provenance of the original data and that of the processes that took place manipulating them should be traceable (Ockeloen et al., 2013; Moreau and Groth, 2013). To facilitate both a bottom-up approach and insight into context and provenance we suggest the following steps:

1) To investigate canonisation, we need to identify *all* names in our datasets and not restrict ourselves to predefined lists. We are, after all, not only looking for the people who made it to the canon, but also for the ones that were forgotten. We therefore need an approach for Named Entity Recognition (NER) to filter out *all* names from our sources. A commonly used state-of-the art named entity recognizer for English reports a 90% F-score (Finkel et al., 2005). However, there are less training sets for Dutch and the task we need in this step is easier than the typical NER task: we are producing lists of people names for historians to study. We therefore mainly need very high recall on identifying person names. Precision is less important, because historians can simply discard expressions that do not refer to a person in their final analysis. Furthermore, we are not interested in names that do not refer to people and standard NER approaches are trained to identify locations, organisations and miscellaneous names in addition to names of people. The exact method we followed for this paper is described in Section 6.
2) Initially, *all* names should be considered as belonging to unique individuals and we should assign all of them an Internationalized Resource Identifier (IRI).[16] We cannot simply assume that the same name refers to the same person. By assigning all names unique IRIs to start with there is no risk of polluting the original data. Any errors can always be traced back to the original source this way (de Boer et al., 2014).
3) The third step is to disambiguate all the names and establish which can be linked to the same person. It is not trivial to do this automatically,[17] but it can be done (as by Veres (Bohannon, 2011)) by comparing the mentioned dates, places, other people and professions in the context. Ideally, the probability of each match should also be indicated. The role of the historian is vital in writing an algorithm for this task, to provide the historical context and establish what can be considered evidence for a match between two people.
4) Most efforts in digitising data evolve around specific 'canonised' topics. We therefore need a non-digitised control dataset to establish in what way the fact that we can only use digitised sources for computational analyses influences the results. For this, a historian still needs to go through the archives to analyse non-digitised sources and write down the names and generic data like dates of birth and death and 'claim to fame'. Of course the historian will once again have to take into account the different kind of sources as mentioned in Section 3. This set should be analysed both apart from and together with the digital set.
5) We would then be able to draw up graphs and tables of which people were mentioned often in what works, when, where and how, which would provide insight in the canonisation of Dutch history.
6) Finally, a more detailed survey should be done by the historian. The leads provided by technology should be followed to see the context and find explanations for the findings. We need access to provenance and context to give room for theory and to assess the meaning of all these numbers (see Hall (2012) for a similar argument).
For this paper we performed step 1 and applied a basic approach to address step 3.

## 6 Results

### 6.1 Top down approach

In this section we will discuss the results from a top down approach for investigating who is most famous in Dutch history. Since any existing fame list would do as a starting point, we took the top 25 of the Dutch TV elections of the 'Grootste Nederlander' (Grandest Dutch person).[18] We then ranked them basing ourselves on the Google (books) Ngram viewer (for English), the KB Ngram viewer, taking the words from Dutch newspapers, the DBNL Ngram

[13] https://books.google.com/ngrams/info

[14] http://www.let.rug.nl/gosse/bin/Web1T5_freq.perl and https://catalog.ldc.upenn.edu/LDC2009T25

[15] http://www.dbnl.org/zoek/ngram.php

[16] IRIs are generalizations of URIs that support Unicode.

[17] Note that there is no predefined ontology, which makes this a different task from standard named entity disambiguation as in (Mendes et al., 2011).

[18] http://nl.wikipedia.org/wiki/De_grootste_Nederlander

viewer, containing words from mostly literary texts and Google Ngrams for Dutch, which contains all words used on the Internet at the end of 2008. With these sets we have sources from historiography, the news, cultural texts and the Internet, which together should provide a rather balanced set of sources with much and less selection, both from the period during and after individuals' lives. We ranked the individuals by their highest score in one year, since for the limited scope of this paper it would go too far to calculate a balanced average for each individual.

We faced several challenges in identifying the right people. The spelling of names is possibly the biggest issue here. Before the nineteenth century there was no standardised spelling of names, which results in many varieties in not only contemporary sources, but also in modern works. Even if a particular name is usually spelled the same way, a bad OCR quality could still give a bias in the results. The options to use wildcards in the viewers to catch all variations often are very limited.

Another problem is caused by people with the same name. William the Silent, number two in the elections (see Table 1), is most commonly known as *William of Orange*. The hits we receive for 'William of Orange' in the Google Ngram viewer however, may refer to the leader of the Dutch revolt (†1584) we are looking for, but also to his great-grandson, the later King of England (†1702), number 72 in the TV elections. Pollution with instances of the king of England could be especially significant in the English corpus of Google books. We therefore only used his nickname 'William the Silent' in this corpus. Despite the significant reduction in hits, he still ranks number 1 in Google books, which further justified our decision.

Identifying the humanist scholar Desiderius Erasmus poses a problem because he is known as 'Erasmus'. Dropping his first name would lead to many additional hits from other people and Google NGrams for Dutch does not even facilitate searching for unigrams. The same applies to the philosopher Baruch de Spinoza. A quick search in the World Biographical Information System[19] shows us that while there are 789 hits for Erasmus, there are 'only' 8 hits for Spinoza (and most of them refer to the correct and the same person) indicating that the risk of pollution is lower. Still, results in Google Ngrams seem significantly inflated for the unigram Spinoza, giving him an extremely high score in 1883. The year 1883 does not have a high score when searching for bigrams of 'Baruch Spinoza', or trigrams of 'Baruch de Spinoza', which strongly suggests that too much pollution occurs when the first name is dropped. We therefore added the results for 'Baruch Spinoza' and 'Baruch de Spinoza', whilst knowing the score does not reflect all references to him.

There also are people who are known differently during their lives, such as members of the royalty. We had to search for both princess and queen Juliana and princess and queen Wilhelmina to obtain the best result. For widely known people like them this problem can be circumvented quite easily, but in other cases specific domain knowledge is needed to find all instances. To give just one example: Dutch treasurer Vincent Cornelisz from the first half of the sixteenth century was very famous in his time, but is currently unknown to a wide audience. In history books he is not only referred to as *Vincent Cornelisz*, but also as *Vincent van Mierop* (a name which was used for the first time by his son, not by him), or as *Vincent Cornelisz van Mierop*. In records of his own time, he was so well known that often he was simply referred to as *master Vincent*, which ironically means that the fame in his own time causes a problem in tracing his fame in our time (ter Braake, 2007, p. 375).

[19] `http://db.saur.de/WBIS/basicSearch.jsf`
The system hosts biographies on 6 million people from 58 biographical archives all over the world.

| Rank | Elections 2004 | Total NGram viewers |
|---|---|---|
| 1 | Pim Fortuijn | Koningin Wilhelmina |
| 2 | Willem van Oranje | **Willem van Oranje** |
| 3 | Willem Drees | Koningin Juliana |
| 4 | Antoni van Leeuwenhoek | **Vincent van Gogh** |
| 5 | Desiderius Erasmus | **Rembrandt van Rijn** |
| 6 | Johan Cruijff | **Anne Frank** |
| 7 | Michiel de Ruyter | Johan Thorbecke |
| 8 | Anne Frank | Christiaan Huygens |
| 9 | Rembrandt van Rijn | **Desiderius Erasmus** |
| 10 | Vincent van Gogh | Prins Claus |

Table 1: The most famous Dutch people in history according to the 2004 TV elections and the Ngram viewers in the tables below

In Tables 1, 2, and 3, we see the top ten occurrences of famous people when searching for the original TV elections top 25. The highest average position in all Ngram viewers is listed in the right column of Table 1. It is very clear that the fame of a person depends greatly on the kind of medium that is used. Number 1 of the TV elections, the politician Pim Fortuijn, only features in the Ngrams for Dutch, which is not surprising since the other lists are for the years 1800-2000 and he only rose to fame in the twenty-first century. Queen Wilhelmina, the number 1 in the Total Ngrams list surprisingly did not make it to the top 10 of the elections. The same can be said for the other members of the royalty, prince Claus and queen Juliana. Apparently they were and are very famous, but are not considered of too much historical significance by the Dutch people. Prime minister Thorbecke claims a high position in the overall ranking due to the many mentions in Dutch newspapers in the middle of the nineteenth century. Christiaan Huygens owes his position primarily to the fact that the DBNL has many of his private letters in its collection. Dutch soccer player Marco van Basten does not make it to the overall top ten, but does score highly in the newspapers and on the Internet. William of Orange/the Silent and painter Vincent van Gogh are the only people who feature in every list. If anything, these tables show how relative fame is. The more (heterogeneous) big datasets we have at our disposal the more balanced the picture will become. In the following subsection we will explore what happens when we use a bottom up approach and try to find the famous people that do not feature on any preexisting list.

| Rank | Google Ngram viewer | KB Ngram viewer |
|---|---|---|
| 1 | William of Orange | Johan Thorbecke |
| 2 | Anne Frank | Koningin Juliana |
| 3 | Koningin Wilhelmina | Koningin Wilhelmina |
| 4 | Vincent van Goghh | Prins Claus |
| 5 | Johan Thorbecke | William of Orange |
| 6 | Antoni van Leeuwenhoek | Rembrandt van Rijn |
| 7 | Koningin Juliana | Vincent van Gogh |
| 8 | Christiaan Huygens | Johan van Oldenbarneveldt |
| 9 | Desiderius Erasmus | Marco van Basten |
| 10 | Rembrandt van Rijn | Desiderius Erasmus |

Table 2: The most famous Dutch people in history according to Google Ngram viewer (1800-2000) and KB Ngram viewer (1800-2000)

| Rank | DBNL Ngram viewer | Google Ngrams for Dutch |
|---|---|---|
| 1 | Christiaan Huygens | Marco van Basten |
| 2 | Rembrandt van Rijn | Anne Frank |
| 3 | Johan Thorbecke | Pim Fortuijn |
| 4 | Desiderius Erasmus | Koningin Wilhelmina |
| 5 | William of Orange | Johan Cruijff |
| 6 | Baruch de Spinoza | Toon Hermans |
| 7 | Koningin Wilhelmina | Koningin Juliana |
| 8 | Willem Drees | Willem van Oranje |
| 9 | Vincent van Gogh | Vincent van Gogh |
| 10 | Johan van Oldenbarneveldt | Prins Claus |

Table 3: The most famous Dutch people in history according to DNBL Ngram viewer (1800-2000) and Google Ngrams for Dutch (2008)

### 6.2 Bottom Up Approach

As mentioned in Section 5, it is relatively easy to identify names with tools for Named Entity Recognition. For this particular study, we use a highly simplistic but effective pattern-matching approach. We select combinations of words that start with a word that starts with a capital (e.g. Willem) and end with a word that starts with a capital (e.g. Oranje), which works fine for Dutch (but would be quite useless for German that capitalises all nouns). Because both the first and last word must start with a capital letter, we avoid the inclusion of words that start the sentence.[20] The algorithm allows for two sequential lower case words within the name, since it is customary to write prepositions and determiners in Dutch names in lower case when they are preceded by a first name or initials. The algorithm can thus capture names such as *Johan Derk van der Capellen tot den Pol*, but no names where three lowercase words follow each other which are extremely rare in Dutch.

For our particular use case, we primarily aim for recall, because (1) historians can immediately filter out the invalid patterns found by our approach and (2) bad patterns are often singletons in the corpus having no or little influence on the top and middle of our frequency based lists. For these reasons, precision can be as low as 5% or 10% and the approach will still serve its purpose (though higher precision does make the historian's job easier). Our basic pattern-matching approach is thus preferable for this particular research over the more sophisticated machine-learning approaches that have higher precision, but lower recall.

We tested our method on the data of the Biography Portal of the Netherlands, an aggregated dataset of 23 different sources, all with their own limitations and biases.[21] A biographical dictionary in itself is a 'canon' of noteworthy people and will therefore not reveal many 'forgotten' people. The Portal nevertheless provides a suitable dataset to try out our methodology. It provides a large volume of descriptive texts in Dutch and the output of our algorithm will reveal what person names occur most in these texts (in their own and in other people's biographies) and thereby showing us a measure of fame after all. The principle of applying our method does not differ from applying it to a set that did not apply any form of selection.

With our approach we could easily get the number of occurrences of entries such as Willem van Oranje (William of Orange). The fact that we also got results from 'Tweede Kamer' (Dutch parliament, 882 hits) 'Den Haag' (The Hague, 830 hits) and 'Staten van Holland' (States of Holland, 420 hits) shows an interesting overall bias towards political history, but can be easily discarded for our purpose here. You do not need to be a domain expert to easily see that these expressions do not refer to people.

Named entity disambiguation is more problematic. After discarding the false hits we have Willem I, Willem II, Willem III, Willem IV and Willem V ranking in the top 10 of our list, but unfortunately there have been many counts, dukes and stadtholders over the centuries who go by that name and title. A problem of a different nature is that we have Willem I, Willem van Oranje and the prins van Oranje ranking high, which could all refer to the same person: William the Silent (of Orange), the number 2 from the TV elections and the overall ranking in Table 1. Hits such as 'Van den Bergh' also causes identity problems, since without the context we cannot see which Van den Bergh this is, or even if he or she is an actual historical person or just a historian who is cited often. Some of the results are quite telling, however. We are quite sure that 'Karel V' will almost always refer to emperor Charles V (and perhaps a few times to the fourteenth century French King) and that Frederik Hendrik and prins Maurits refer to the famous sons of William the Silent. Domela Nieuwenhuis must refer to the social anarchist Ferdinand Domela Nieuwenhuis, since he has quite a unique name.

In a first attempt of named entity disambiguation we investigated the possibilities of applying time constraints based on metadata and temporal expressions in the text. This way count Willem II (thirteenth century), stadtholder Willem II (seventeenth century) and king Willem II (nineteenth century) would be easily separated.

We implemented a basic approach that tackles the time constraint of identity, which is based on the idea that people can only personally interact with someone who was alive at the same time as they were. Because this is the case, we as-

[20] Names such as *Willem II* are identified, because the sources use Roman capital letters to add numbers to nobility with the same first name.

[21] http://www.biografischportaal.nl

| Rank | BP first results | BP second results |
|---|---|---|
| 1 | Willem I | Willem I (1772) |
| 2 | Willem III | Karel V (1500) |
| 3 | Prins van Oranje | Willem II (1792) |
| 4 | Karel V | Willem III (1650) |
| 5 | Willem II | Willem V (1748) |
| 6 | Willem V | Domela Nieuwenhuis (1846) |
| 7 | Frederik Hendrik | Frederik Hendrik (1584) |
| 8 | Domela Nieuwenhuis | Willem III (1817) |
| 9 | Willem IV | Lodewijk Napoleon (1778) |
| 10 | Prins Maurits | Willem IV (1711) |

Table 4: Results from the Biography Portal of the Netherlands, without (left) and with (right) time disambiguation. The second column also shows year of birth

sume that in the typical case, people who are mentioned in someone's biography will be a contemporary of the biography's subject. In order to establish which mentions refer to the same person, we extracted the date of birth and date of death from the metadata of the biographies in our corpus. While going through the corpus to identify names, we only merged names when the lifespan of the subjects either overlapped or were maximum 50 years apart from each other. This baseline assures that, if the reference in the text itself is not about the far past or future, it is at least possible that the texts refer to the same person.

Because there may be people alive at the same time who have the same name and 50 years offers quite a range, the approach does not offer any guarantees that references to different people are not combined, but it helps to solve some of the clearer cases where sources do not talk about the same person. It solves, for instance, the issue of high nobility with the exact same name. They are either from a different era altogether, or they have a different number behind their name.

The results of this approach are quite promising. Table 4 shows that while we previously were not able to distinguish between Willem III, the nineteenth century king and Willem III the seventeenth century stadtholder, we now have them listed as two different individuals. It also shows that Willem I does not refer to William the Silent at all, as one may expect from the lists from our top down approach, but to nineteenth century king Willem I. Looking at the tables it seems that the Biography Portal of the Netherlands, and then most likely especially the two biggest dictionaries included in there from the nineteenth and early twentieth century, are strongly biased towards the House of Orange. Further research might show that many people were included in the dictionaries *because* of their link to king Willem I.

By refining this method, for example by automatically merging similar instances like 'Willem I' and 'Koning Willem I', and by applying it to a larger and a wider variety of datasets we would become closer to seeing canonisation patterns than traditional research could have ever brought us.

In an attempt to trace the 'forgotten' individuals we made a list of the people who do get mentioned frequently in the texts from the BP, but who do not have a biographical entry of their own. The results of this exercise were interesting enough, but do still involve quite a lot of handwork from the historian. Many people in the list we generated did have their own entry after all, but are mentioned in a slightly different way. Politician P.W.A. Cort van der Linden, for example, is often mentioned as *Cort van der Linden* (16 times) and similar issues occur with many other politicians from the nineteenth and twentieth century. Moreover, some individuals are known under various alternative names. For instance, sixteenth century duke Karel van Gelre is listed as *Karel van Egmond*.

The people who are mentioned most frequently in the texts and who really *do not* have their own biographical entry are listed in Table 5. We find an important religious figure, a communist philosopher (probably mainly thanks to the biographical dictionary on socialists included in the Portal), no less than eight French rulers, an English king and a German emperor in the top 12. It does not bring us closer to the forgotten people in Dutch history, but does show a clear connection of Dutch elites with French royalty (or a bias in the dictionaries towards France or people involved with France). We also encounter the previously mentioned problem of how to identify people who are mostly known with one name. To detect Erasmus in the Ngram viewers we had to search for Desiderius Erasmus. In Table 5 we see 15 mentions of Napoleon Bonaparte, while there will be many more for just Napoleon. To find them, however, we would have to expand our algorithm to include one word instances as well, which would result in too much noise for our analyses for this basic version of our algorithm.

To trace the individuals who were noteworthy in their own time, but are forgotten in history, we are more likely to be successful when analysing sources with a semi-conscious selection mentioned in Figure 1. We applied our method to a sample of 99 historic Dutch newspaper texts provided by the Koninklijke Biobliotheek.[22] The sample is too small to provide indications of 'forgotten people', but the outcome of this test shows that our method can be applied successfully to these articles. The outcome furthermore confirmed our observation based on data from the BP that phrases that do not correspond to a name generally occur only once and therefore do not form a hindrance for the historian, given that a single mention does not point to (contemporary) fame.

## 7 Conclusions

In this paper we addressed the importance of research on canon formation in historical research. Before the advent of digital technologies and the availability of digitised data, this could only be done tentatively. We have shown that despite many methodological and technical problems, there is a decent amount of data available and there are tools that facilitate group analyses of famous people.

In section 5, we proposed a method to complement a top-down approach of analysing people still famous now with a bottom-up approach, which gives more room for unbiased selection, context and provenance of the data. The basic

[22] http://lab.kbresearch.nl/get/Downloads

| Rank | Individuals without their own biography | Number of mentions |
|---|---|---|
| 1 | Jezus Christus | $> 75$ |
| 2 | Karel II (king of England) | 60 |
| 3 | Lodewijk XIV (king of France) | 40 |
| 4 | Lodewijk VIII (king of France) | 25 |
| 5 | Lodewijk XI (king of France) | 25 |
| 6 | Frans I (king of France) | 23 |
| 7 | Lodewijk XII (king of France) | 18 |
| 8 | Karl Marx (German philosopher) | 18 |
| 9 | Lodewijk XVI ((king of France) | 16 |
| 10 | Jozef II (German emperor) | 15 |
| 11 | Lodewijk XIII (king of France) | 15 |
| 12 | Napoleon Bonaparte (French emperor) | 15 |

Table 5: People mentioned most frequently in the Biography Portal of the Netherlands, without their own biographical entry

means to carry out such research are available. Even though methodologies for task 3) are still in a preliminary stage and the work in 4) and 5) still is labor intensive, the possibilities provided by digital humanities make this research feasible.

In section 6, we discussed some difficulties in applying a top-down approach and have also discussed the first results of a bottom-up approach. A close collaboration between historians and computer scientists is a requirement to make such research successful, especially in the named entity disambiguation. Expert domain knowledge combined with complex algorithms are needed to match as many individuals correctly as possible and to signal false positives. Eventually such exercises can help us to explain why some people only get 15 minutes of fame and others live on in memory over centuries.

The approaches we presented in this paper are relatively basic. We explained that this is not an issue for named entity recognition, because precision is of minor importance for the historian investigating canonisation. We plan to experiment with alternative versions of the algorithm including a version that can handle single names such as Erasmus and Napoleon. However, given that our basic algorithm already provides results that yield interesting results, future work will mainly focus on better disambiguation. We expect that standard methods for named entity disambiguation are not the most suitable for this task and data, because they tend to make use of the content words used in the text and address a wider range of named entities than just people. We therefore expect most from a domain and target entity specific approach that combines frequency of the first and last name, information about time and place, as well as social networks.

The most important next step, however, will be to apply the methods outlined in this paper to new datasets that also provide a contemporary perspective and/or use semi-conscious selection. Contemporary sources play a vital role in identifying people who were famous and fell in oblivion thus providing the necessary means to compare and identify what aspects contribute to canonisation once initial fame is achieved.

## 8 Acknowledgements

This work was supported by the BiographyNet project `http://www.biographynet.nl` (Nr. 660.011.308), funded by the Netherlands eScience Center (http://esciencecenter.nl/). Partners in this project are the Netherlands eScience Center, the Huygens/ING Institute of the Royal Dutch Academy of Sciences and VU University Amsterdam. We would like to thank Dr. Ronald Sluijter for his insightful comments on an earlier version of this paper. All remaining errors are our own.

## 9 References

J. Bohannon. 2011. Google books, wikipedia, and the future of culturomics. *Science*, 131:135.

M. Bosch. 2014. 1001 vrouwen in perspectief. traditie en verandering van het biografische woordenboek in nederland en elders. *BMGN, LCHR*, 129(1):55–76.

V. de Boer, J. Leinenga, M. van Rossum, and R. Hoekstra. 2014. Dutch ships and sailors linked data cloud. In *Proceedings of the International Semantic Web Conference (ISWC 2014), 19-23 October, Riva del Garda, Italy*.

A.E. Earhart. 2012. Can information be unfettered?: Race and the new digital humanities canon. In M. K. Gold, editor, *Debates in the Digital Humanities*, pages 309–318. University of Minnesota Press, Minneapolis, London.

J.R. Finkel, T. Grenager, and Ch. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05). Association for Computational Linguistics, Stroudsburg, PA, USA*.

A.S. Fokkens, S. ter Braake, N. Ockeloen, P. Vossen, S. Legêne, and G. Schreiber. 2014. Biographynet: Methodological issues when nlp supports historical research. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland*, May.

M. Halbwachs. 1985. *Das kollektive Gedchtnis. Mit einem Geleitwort zur deutschen Ausgabe von Heinz Maus*. Fischer, Frankfurt am Main.

G. Hall. 2012. Has critical theory run out of time for data-driven scholarship? In M. K. Gold, editor, *Debates in the Digital Humanities*, pages 127–132. University of Minnesota Press, Minneapolis, London.

L. Hanssen. 1995. Op zoek naar een onbekende. biograsche lexicons als wetenschappelijk hulpmiddel. *Biografisch Bulletin*, 5(1):77–83.

P. N. Mendes, M. Jakob, A. Garca-Silva, and Ch. Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *7th International Conference on Semantic Systems (I-Semantics '11)*.

J.B. Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 131:176–182.

L. Moreau and P. Groth. 2013. *Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool.

I. Nadel. 1984. *Biography. Fiction, fact & form*. MacMillan, London and Basingstoke.

N. Ockeloen, A.S. Fokkens, S. ter Braake, P. Vossen, V. de Boer, G. Schreiber, and S. Legêne. 2013. Biographynet: Managing provenance at multiple levels and from different perspectives. In *Proceedings of the Workshop on Linked Science (LISC2013) at ISWC (2013)*.

R. Rosenzweig. 2011. Wikipedia: Can history be open source? In R. Rosenzweig, editor, *Clio Wired. The Future of the Past in the Digital Age*, pages 51–82. Columbia University Press, New York.

M.L. Sample. 2012. Unseen and unremarked on: Don DeLillo and the failure of the digital humanities. In M. K. Gold, editor, *Debates in the Digital Humanities*, pages 187–201. University of Minnesota Press, Minneapolis, London.

S. ter Braake. 2007. *Met Recht en Rekenschap. De ambtenaren bij het Hof van Holland en de Haagse Rekenkamer in de Habsburgse Tijd (1483-1558)*. Verloren, Hilversum.

M. Wilkens. 2012. Canons, close reading, and the evolution of method. In M. K. Gold, editor, *Debates in the Digital Humanities*, pages 249–258. University of Minnesota Press, Minneapolis, London.