How to Make it in History

Working Towards a Methodology of Canon Research with Digital Methods

Classic Historical Questions:

Why are some people remembered and others forgotten?

What does this tell us about canon formation, memory and group identity?

Difficult to research with traditional means. Necessary to analyze huge groups of people over a longer period of time.

Computational Methods May Help

The Science Hall of Fame

(Veres and Bohannon)

Bertrand Russell	1872	1970	1500
Charles Darwin	1809	1882	1000
Albert Einstein	1879	1955	878
Lewis Carroll	1832	1898	479
Claude Bernard	1813	1878	429

What scientists are mentioned most frequently in the corpus of Google Books?

Method:
- Mining Scientists from Wikipedia and the Encyclopedia Britannica
- Measuring their fame by looking at the frequency they are mentioned
- Delete outliers

Famous Dutch People According to Several N-Gram Viewers. A little test case:

Method:
- Take the top 25 of the 2004 tv elections for the grandest Dutch person.
- Look at how often they are mentioned in Google NGrams (corpus English), Ngrams for Dutch (names mentioned on the internet end 2009), KB Ngram viewer (newspapers), DBNL Ngram viewer (iterary works)
- Rank them by high score and see what happens
- Results:

Google Books	KB News Paper Archive	DBNL Literary Texts	Dutch Ngrams internet 2009	Over all ranking
Willem van Oranje/de Zwijger	Johan (Rudolph) Thorbecke	Christia(a)n Huygens	Marco van Basten	Koningin/prinses Wilhelmina
Anne Frank	Koningin/prinses Juliana	Rembrandt van Rijn	Anne Frank	Willem van Oranje/de Zwijger
Koningin/prinses Wilhelmina	Koningin/prinses Wilhelmina	Johan (Rudolph) Thorbecke	Pim Fortuyn/Fortuijn	Koningin/prinses Juliana
Vincent van Gogh	Prins Claus	Desiderius Erasmus	Koningin/prinses Wilhelmina	Vincent van Gogh
Johan (Rudolph) Thorbecke	Willem van Oranje/de Zwijger	Willem van Oranje/de Zwij-ger	Johan Cruijff/Cruyff	Rembrandt van Rijn
(Antoni) van Leeuwenhoek	Rembrandt van Rijn	Baruch (de) Spinoza	Toon Hermans	Anne Frank
Koningin/prinses Juliana	Vincent van Gogh	Koningin/prinses Wilhelmina	Koningin/prinses Juliana	Johan (Rudolph) Thorbecke
Christia(a)n Huygens	Johan van Oldenbarnevel(d)t	Willem Drees	Willem van Oranje/de Zwij-ger	Christia(a)n Huygens
Desiderius Erasmus	Marco van Basten	Vincent van Gogh	Vincent van Gogh	Desiderius Erasmus
Rembrandt van Rijn	Desiderius Erasmus	Johan van Oldenbarnevel(d)t	Prins Claus	Prins Claus

Methodological Problems

Problems of biases in the corpus

Problems with OCR

Problems with identifying names (different spelling; which Erasmus?)

Most Importantly: Where do the original names come from? In other words: how do we know who to look for? Taking a list of already canonized names might only reaffirm the canon

6 step solution?

1)Named Entity Recognition: Mine large corpora of texts for any names; recall most important; checking words for capital letters might be even more effective; delete outliers (geographical names, names of institutions, important events et cetera)

2) Assign IRI's to all found entities in one dataset

3) Named Entity Disambiguation: When are we speaking about the same people? Domain specific knowledge vital to set the parameters for an algorithm.

4) Manual check by mining people in non-digitized sources. Where are the biases in our corpora? How can we solve this or what does this mean for our computerized analysis?

5) Bring all found people together in one supra dataset and disambiguate once more.

6) Analyze the findings. What patterns in fame are there? What people are likely to be remembered or quickly forgotten?

A New Canon of Dutch History?

Serge ter Braake and Antske Fokkens

netherlands eScience center

VU UNIVERSITY AMSTERDAM