Bias in DH

Bias in DH

Some biases are inherent to the digital methods used in DH.

Some are inherent to the core SSH disciplines themselves.

-> focus on text-based discipline with a historical component: literary studies, and especially literary history

-> then focus on corpus without historical component: twitter corpus analysis

Bias in literary studies

And what is means for the literary canon

Bias in literary history

Biases in literary history:

- Constitution
- Reception
- Digital Approach

Literature History connected to Cultural Heritage -> key issue of Authority.

Collecting, presenting, editing

Collecting and presenting texts = one of the core missions of literary studies.

Central role of editions in presenting authoritative instances of text representations.

Types of editions closely linked to philological traditions (different in German-speaking, French-speaking, English-speaking etc. areas).

Collecting as a core issue

Collecting:

- in literary studies, preliminary work (non authoritative, black box).
- in DH, essential step in the work with data
- in cultural heritage, authoritative even if there is no textual representation of the data (image, metadata)

=> different status - and yet, all deal with the same objects!

Collecting, sorting, selecting

Collecting process supposes to make choices:

- not everything will be kept,
- different periods of time/places have different selection criteria (and yet we are dependent on the criteria of earlier periods/other places),
- there exist different ways of presenting the selected documents/information

Selecting means: choosing, defining criteria for selection, tossing out, in some cases even destroying documents.

How to collect when you know of the biases

Katherine Bode's critical view on Moretti's approach to distant reading: "Rather than just adding women, non-white, working-class, and non-Western authors to the canon, distant reading would incorporate all literature into its analyses."

Katherine Bode, "Why you can't model away bias". Preprint: Modern Language Quarterly 80.3

Convergence between cultural heritage and ML?

Gebru paper “Lessons from Archives”: no particular consideration for literary studies, convergence is between cultural heritage and ML.

Identifies key values in archival/recording work that can at different levels be applied in an exemplary manner to ML:

- Consent (while collecting)
- Inclusivity (mission statement explaining collecting principles)
- Power (building consortia)
- Transparency (keep track of recording process)
- Ethics & Privacy (professional record-keeping)

Lessons from Archives

Data Statement as key to identifying biases: self-reflexion within the collecting process

Elements put forward by Gebru as useful for ML:

- Communication between institutions
- Crowdsourcing
- Documentation

Interest for local initiatives asks the question of the implementation of this convergence at a larger scale

From collecting to establishing a literary canon

Literary history: not only collection in the sense of record-keeping, but also a canon

-> Stanford Literary Lab Pamphlet #11: trying to address how archive and canon relate to one another, the paper moves to a different question.

Convergence of “the published”, “the archive” and “the corpus” in the digital age as a starting point.

“Canon” as the power of those who can select texts (for anthologies for instance); based on the criteria of popularity and prestige

Biases in the digital literary canon

Jokers, Moretti, Bode, all point to the fact that the literary canon as we know it represents only a (biased) fraction of the literary production (even only taking into account the published literature recorded in libraries) -> Difficulty to encompass “all”

Genre of the novel particularly appropriate to track and try to correct biases: see “Sampling criteria” of ELTeC: https://distantreading.github.io/sampling_proposal.html

Among the novels in each language the subcollection must contain...

- at least 10%-50% have been written by female authors for the language subcollection.
- 9 to 11 authors are represented with exact three novels.
- at least 30% are highly canonized novels, at least 30% should be non-canonized novels, based on the following reprint groups: reprinted not at all, reprinted once, reprinted more than once within the period 1980-2000
- at least 20% are short novels (10-50k word tokens), at least 20% are long novels (>100k word tokens).

Example of early 19th century literature

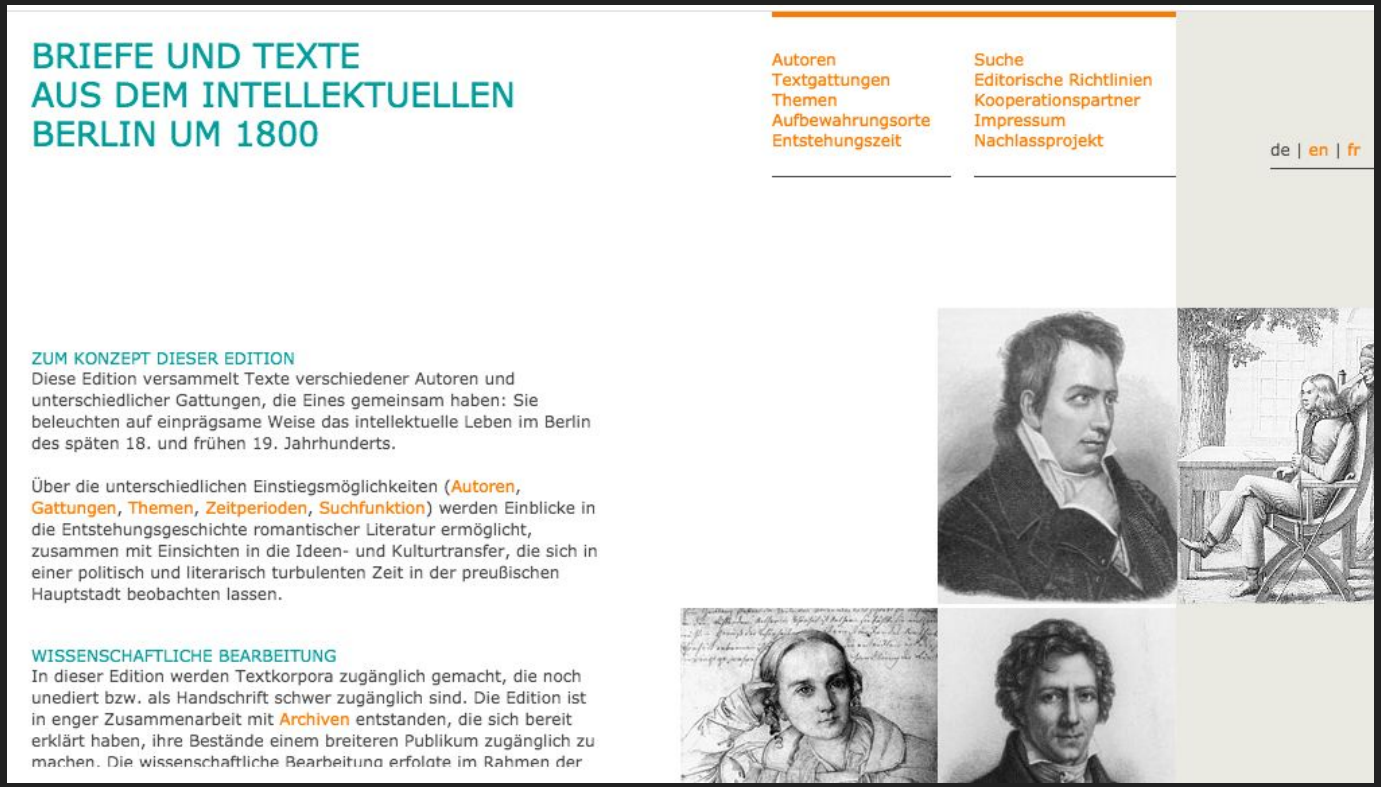
Two major approaches: genres and/or authors

Emergence of “big (male) authors”

Editions of texts (rather works in this case) still defined today either by author or literary genre, but mostly by their uniqueness (first exceptions for romantic circles in the 20th century, but rather seldom)

Underrepresentation of women writers and of the genres they used

Trying to present the diversity of literary production



Avoiding biases?

Conservation & publications have inherent biases, cannot be avoided.

Bode: giving the conditions for accountability (p. 14)

Biases can be identified and reduced during corpus construction by researching precisely the origin of documents & publication context