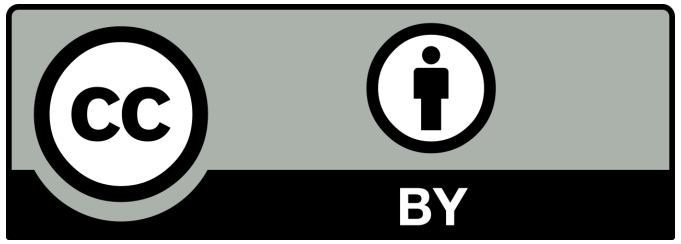
Digital Humanities and Machine Learning – as seen from the DH perspective

Anne Baillot, Le Mans Université

July 2021



Defining Digital Humanities

- One the main activities of Digital Humanists = defining (and/or translating) DH

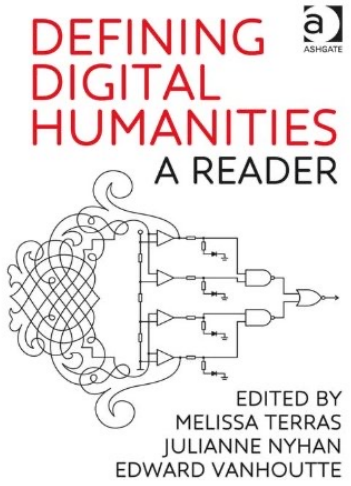
-> Mareike König, *Was sind Digital Humanities? Definitionsfragen und Praxisbeispiele aus der Geschichtswissenschaft*:

https://dhdhi.hypotheses.org/2642

-> *Debates in the DH* « What is Digital Humanities

and what's it doing in English Departments »

by M. Kirschenbaum: https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/f5640d43-b8eb-4d49-bc4b-eb31a16f3d06



- « Digital Humanities »: **contradiction in terms or more information than needed**?

What DH does what traditional Humanities don’t

- Extensive use of computing, quantitative methods, including the development of dedicated digital tools
- Online dissemination of scholarly output
- Epistemological dimension:
  - reflecting on what the extensive use of computers changes for Humanities disciplines
  - reconsidering the disciplinary structure of knowledge established since the 19th century
- For young scholars, disciplinary recognition is key to get funding, reputation, positions…: these questions are not theoretical issues, they are very practical in terms of academic career

DH in Germany and in the world

- Germany: DH curricula, professors, a growing research infrastructure – but no DFG Fachkollegium
- DH communities organized in associations:
  - DHd-Verband for German speaking countries: https://dig-hum.de/
  - EADH at European level: https://eadh.org/
  - ADHO at global level: https://adho.org/
- Each level has their own conferences:
  - DHd 2022 in Potsdam: https://www.dhd2022.de/
  - DH 2023 in Graz
  - DH Stammtisch in Berlin (first Friday of each month)

A few DH projects from Germany & German-speaking countries

- Drama Corpora Project DRACOR: https://dracor.org/
- European Literary Text Collection ELTeC: https://www.distant-reading.net/eltec/
- @Rotrechnen for history paintings: https://data.ub.uni-muenchen.de/82/1/Rotspektralanalysen_Herrscher_und_Politikerbilder_01.pdf
- Sentiment analysis at state level (based on documents from Swiss diplomacy): https://sciendo.com/article/10.2478/ADHI-2018-0044
- Distant reading in contemporary history (based on GDR press): https://zeithistorische-forschungen.de/1-2019/5694

What kind of research questions are relevant?

- DH as an opportunity to answer « old » Humanities questions (e.g., anonymous authors who can be identified with stylo)
- DH as a field of possibilities; new research questions adressable (for instance by harvesting documents scattered around the globe, entailing new forms of aggregation and comparison)
- Multi-layered projects: several aspects relevant for research including methodology, analysis, and interpretation

=> challenge: find research questions that will be as interesting for ML as they are for DH (and Humanities disciplines and data management)

Text in text-based DH

- Notion of what is « a good text » => reflected in digital methods:
  - corpus-based
  - raw text vs. annotated text
  - philological requirements (critical editions)
- Web 2.0 adds potential of connecting information between texts via the internet
- Development of semantic web and ontologies in the past decade
- Since the 1990s, development of an XML-based specification for the annotation of text: the Text Encoding Initiative, now a standard.

Project lifecycle

- Define research questions (DH+ML)
- Set up method and timeline, including iterations (gathering data, annotation, training, testing, evaluation, interpretation)

$\Rightarrow$this step takes A LOT of time due to the different scholarly cultures, disciplines, vocabularies involved (see for instance the endless debates on what a model is)

- Every step of the project is to be documented thoroughly; ideally each data set and script collection should be made availabe for reuse, reproduction and verification purposes (not always possible, for instance in the case of tweets or googlebooks scans)

Originality of such a project lifecycle, as seen from the Humanities' side

- Involves a team with different competences
- Allows to tackle much larger research issues than what is usually addressed in the Humanities (in terms of data quantity, of disciplinary orientation)
- Questions of copyright, publication rights, privacy are likely to arise when using data that is gathered/(made) available online
- Publication output much more varied and rich than traditional Humanities papers/books (datasets, scripts, software, online tools, documentation…)

=> still a strong inhibition towards such an approach in the traditional Humanities

DH as an interface between ML and Humanities

- DH community actively involved in opening up Humanities research to what is happening online at large
- Strong advocacy for digital infrastructures, formats (standards), legal issues, Open Science within the DH community
- Specific reputation mechanisms (in general, two disciplinary fields to be satisfied => massive research output in order to get the scholarly recognition that is necessary in both fields)
- Divide with ML community on some points (pre-print culture, role of conference publications, h-index)

Conceiving a DH project that is actually also interesting for ML

Conceiving a DH project that is actually also interesting for ML

To what extent do egodocuments present an emotionality of hatred similar to that of official discourses in the context of French-German relationships, taking into account their evolution in times of war and in times of peace since the end of the 18th century?

Building block 1: egodocuments

- Corpus construction: choice of egodocuments
- Written in the first person, subjective view on events (everyday life, political events)
- Not necessarily aiming at a publication: concerns high and low literature
- Egodocuments: key for History (history from below) and literary studies
- And yet no large scale studies on egodocuments that involve wide corpus and variety of scriptors => explorative work beginning with construction of corpus

Building block 2: historical width

- Corpus construction: choice of period of time considered
- Variety and quantity of data requires a long period of time: 18th-20th century.
- But: languages evolve along time and scriptor profiles change over time (need to be literate & wealthy especially in the pre-modern era) => several language models needed for the ML approach
- Preservation and access issues impact representativity in corpus construction
- Raises questions re:digitization of cultural heritage: what should be made available and reusable? How to make it findable? How to present it in such a way that it can be used by research? How to make sure corpora are not biased?

Building block 3: German-French comparison

- Twofold comparison: German-French and War-Peace (comparison=>reducing biases? will need to be verified)
- Comparison based on socio-economic similarities of the 2 countries with focus on Franco-German wars and narratives of the enemy
- Corpus annotation: creating categories that are first applied manually on part of the corpus
- Challenge: training a machine to apply tags on texts in different languages, from different periods and contexts. Question: is translation helpful (in order to have one language of reference)?=> will have to be assessed.

Building block 4: sentiment analysis

- Manually annotated corpus is used for training on another part of the corpus (yet another part has to be reserved for testing).
- Goal: defining sentiment toward the enemy (French for German and German for French) and its evolution over time.
- Challenge: go beyond positive/neutral/negative, analyze nuances and their evolution over time
- Reference corpus for the hatred rhetorics of each period: press, political discourse

=> goal: observe the type of connection between emotivity and the evolution of first-person writing

Expected Output

- Literary studies:
  - full edition of the first-ever corpus of egodocuments that large
  - questioning literary history via ordinary writings (what is high/what is low literature)
- History:
  - first-ever documentation and annotation of testimonies on a wide range of periods based on thematic focus
  - novel mixed approach of political history and mentality history
- ML:
  - development and training of new language models
  - improving methods for limited data amount (with several language models)
  - improving debiasing methods in a historical perspective (with several language models)
  - improving sentiment analysis state of the art (with several language models)

The questions I want answers to

- Did German and French actually hate each other or not? Does writing about one's life change anything about it?
- Did German and French hate each other (as a nation) more in war times or are the feelings concentrated on specific individuals/social classes?
- Do the feelings towards the enemy evolve in the same way in France and in Germany in the two past centuries?
- How do ideas developed by a small circle of intellectuals actually reach « normal people » to the point where they identify themselves with these ideas?

Reuse, verify, reproduce

- Added value of such a DH project: allowing further research (corpus constitution, methods, and their evaluation)
- Output must be reusable: findable, accessible, enrichable
- Documentation is key! And better yet: a data management plan.
- Reproducibility: debate in the Humanities, key in DH. Works with explicitation of analysis work step between corpus constitution and interpretation.
- Choice of trusted repositories: see OpenDOAR

Publication strategies in DH

- Publications in DH only slowly evolving towards dynamic formats allowing more diverse forms of results representation (e.g., Jupyter notebooks)
- Strong feelings about citability of online resources in a world still dominated by page numbers
- Development of preprint-based overlay journals
- Older journals (DHQ, DSH), and newer ones: ZfdG, JDMDH, JCLS

Stay sober!

- Environmental footprint of computing-intensive research: machines, data flow, data storage,…

-> minimal computing: https://go-dh.github.io/mincomp/about/

- Dissemination strategies and mobility also concerned
- Environmental footprint = CO2 production, impact on biodiversity, impact on resources, esp. water resources

=> need to advocate for a greater transparency from research infrastructures, and ways to calculate the environmental impact of research activities easily

And now…

have fun!