Humanities in the Digital Age

University of Warsaw, Instytut Informatyki,
May 18-20, 2017
AI writing

*The Drum* – the first “driverless magazine”

2016

Welcome to the digital age (in 1833)

The invention of digital communication in 1833
Carl Friedrich Gauss and Wilhelm Weber at Göttingen
Data Revolution

Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time, designing, monitoring and evaluating effective policies becomes almost impossible. New technologies are leading to an exponential increase in the volume and types of data available, creating unprecedented possibilities for informing and transforming society and protecting the environment. Governments, companies, researchers and citizen groups are in a ferment of experimentation, innovation and adaptation to the new world of data, a world in which data are bigger, faster and more detailed than ever before. This is the data revolution.

UN Data Revolution Report “A world that counts”, 2015
http://www.undatarevolution.org/report/
I. The rise of computing in the humanities
1940 Roberto Busa SJ working on his PhD thesis about *praesentia* in the work of St. Thomas
   The role of function words to understand concepts: *in praesentia*
   The solution: 10,000 file cards

Humanities, going digital
1946 Roberto Busa SJ building a concordance of all words used by Thomas Acquinas

1949 ....or punchcards?
Index Thomisticus

1949
Roberto Busa SJ (1913-2011) &
Thomas J. Watson (Gründer von IBM)

56 volumes
http://www.corpusthomisticum.org
Tool or method?

Massive work, done by computers -> computer as a tool (e.g. bibliographies)

Roberto Busa: A new approach using function words instead of meaningful words -> computer as a method

P. Busa SJ at the world expo, Brussels, 1958
Federalist Papers

Series of 85 articles, first published in a variety of newspapers in 1787/88 under the name of „Philo-Publius“ (i.e. Alexander Hamilton, James Madison, John Jay)

1788 published as „The Federalists. A Collection of Essays“

1940 and 1964 probabilistic approach: „Inference and Disputed Authorship. Federalist Papers“ by Frederick Mosteller and David L. Wallace
Morelli’s methods

Ivan Lermolieff [i.e. Giovanni Morelli]
Ein kritischer Versuch — Leipzig, 1880,
URL: http://digi.ub.uni-heidelberg.de/diglit/morelli1880/0116
Authorship attribution

Style, the unconscious use of words, mainly little words – a kind of a fingerprint to detect authorship
St. Paul and the riddle of authorship

Average of word length (measured in syllables)

Augustus de Morgan (1851), Letters


Who wrote Shakespeare’s plays?

„it is proposed to analyze a composition by forming what may be called a ‘word spectrum’ or ‘characteristic curve’, which shall be a graphic representation of the arrangement of words according to their length and to the relative frequency of their occurrence. [...] the method] may equally well be applied to the study of syllables, of words in sentences, and in various other ways“

Thomas Corwin Mendenhall
Laws of stylometry

Wincenty Lutosławski (1898),
Principes de stylométrie,
Revue des études grécques 41, 61-81

Loi stylométrique:
“De deux échantillons de texte de la même longueur comparés avec un troisième sous le rapport du style, celui qui présente une affinité stylistique décidément plus grande avec l’étalon de comparaison, lui sera plus proche quant à la date de la composition, pourvu qu’un nombre suffisant de stylèmes ait été étudié et inclus dans la calculation des affinités.”

Formalism

Figure 4 shows how the combinations of the six terms, under two different

![Diagram showing combinations of six terms]

FIGURE 4. Pottery, main body. 9 elementary profiles.

![Diagram showing examples of code words for different shapes]

FIGURE 5. Pottery, main body. Examples of code words for different shapes.

Jean-Claude Gardin,
- Le Fichier mécanographique de l'outillage: outils en métal de l'âge du bronze, des Balkans à l'Indus, 1956
- On the Coding of Geometrical Shapes And Other Representations, with Reference to Archaeological Documents, 1959
Laws of story

31 generic functions/narratemes

The combination of moves may be as follows:

1. One move directly follows another. An approximate scheme of such combinations is:

   I. A ———— W*

   II. A ———— W2

2. A new move begins before the termination of the first one. Action is interrupted by an episodic move. After the completion of the episode, the completion of the first move follows as well. The scheme is:

   I. A ———— G ———— K ———— W*

   II. a ———— K

3. An episode may also be interrupted in its turn, and in this case fairly complicated schemes may result.

   I. ———— G ———— K ———— W*

   II. ———— W*

   III. ———— K

4. A tale may begin with two villainies at once, of which the first one may be liquidated completely before the other is. If the hero is killed and a magical agent is stolen from him, then first of all the murder is liquidated, and then the theft is liquidated also.

   A214{ I. ———— K9

   II. ———— K1

5. Two moves may have a common ending.

   I. ———— W*

   II.
Principles and Patterns

Rens Bod,
A New History of the Humanities.
The Search for Principles and Patterns from Antiquity to the Present,
2013
Cultural heritage

»Damage to cultural property, belonging to any people whatsoever, means damage to the cultural heritage of all mankind, since each person makes its contribution to the culture of the world.«

The Hague Convention, 1954

Giovanni Pannini, Roman Ruins with a Prophet, 1751
The works touching books are two — first, libraries, which are as the shrines where all the relics of the ancient saints, full of true virtue, and that without delusion or imposture, are preserved and reposed; secondly, new editions of authors, with more correct impressions, more faithful translations, more profitable glosses, more diligent annotations, and the like.

Corpus Inscriptionum Latinarum

Theodor Mommsen
1853 ff.
17 volumes, in 70 parts
ca. 180,000 Inscriptions

http://cil.bbaw.de/
Grimm’s dictionary

http://dwb.uni-trier.de/de/
Scholarly Primitives

- Discovering / Retrieval
- Annotating
- Comparing
- Referring
- Sampling
- Illustrating
- Representing

John Unsworth (2000), Scholarly Primitives:

What methods do humanities researchers have in common, and how might our tools reflect this?
Historical and hermeneutical approaches

Study room of Jacob Grimm
Digital scholarship

- Library catalogues and scans
- Digital editions, bibliographies, and dictionaries
- Wikipedia
- Google Books

and/or

- Big data and large corpora
- High precision
- Better sharing of results
- New approaches and findings
Two cultures

Charles P. Snow, *The two cultures*, 1959
Digital Humanities – Third Culture?

- ‘Humanities computing’
- ‘Digital Humanities’
New Paradigm?

Tony Hey (2009), The Fourth Paradigm
Chris Anderson (2008), The End of Theory
Science Special (2011), Dealing with Data
Library Offers Largest Release of Digital Catalog Records in History

25 Million Free Records of Bibliographic Metadata

Press Contact: Shenon Connady (202) 707-0450

The Library of Congress announced today that it is making 25 million records in its online catalog available for free bulk download at loc.gov/cds/products/marcDist.php. This is the largest release of digital records in the Library’s history.

The records also will be easily accessible at data.gov, the open-government website hosted by the General Services Administration (GSA). Until now, these bibliographic records have only been available individually or through a paid subscription.

The Library also is joining with George Washington University and George Mason University to host a Hack-to-Learn workshop Wednesday, May 17 through Thursday, May 18, which will bring together librarians, digital researchers and coders to explore how the data (and other interesting data sets) can be used.

“The Library of Congress is our nation’s monument to knowledge and we need to make sure the doors are open wide for everyone, not just physically but digitally too,” said Librarian of Congress Carla Hayden. “Unlocking the rich data in the Library’s online catalog is a great step forward. I’m excited to see how people will put this information to use.”

The new, free service will operate in parallel with the Library’s fee-based MARC Distribution Service, which is used extensively by large commercial customers and libraries. All records use the MARC (Machine-Readable Cataloging Records) format, which is the international standard maintained by the Library of Congress with participation and support of libraries and librarians worldwide for the representation and communication of bibliographic and related information in machine-readable form.

The data covers a wide range of Library items including books, serials, computer files, manuscripts, maps,
According to Google more than 130 million different books were published, 30-40 mio. digitized by Google

Google Books, Gutenberg, Early English Books online, ECCO, Evans, British National Corpus, Gallica, VD 16. / VD 17./ VD 18., TextGrid, Deutsches Textarchiv (DTA), Zeno.org ...
Digital Humanities - normal science or nerds?
Digital Humanities

- One area of computer science (like computational chemistry etc.)
- One form of the humanities
II. Humanities computing: Cultural heritage
Universal Leonardo

http://www.universalleonardo.org
Digital archaeology

Angkor Wat

Light detection and ranging (Lidar)
Theodor Fontane, Notebooks
<table>
<thead>
<tr>
<th>HOME</th>
<th>Sammlungsobjekte</th>
<th>Literatur</th>
<th>Volltexte</th>
<th>Bibliographie</th>
</tr>
</thead>
</table>

**Tahitische „Malayische“**
1794, Schädel-Decaden III (1795, S. 48-47) Dissertation Blumenbach (deutsche Ausgabe, 1798, Verzeichnis, S. 3)

**Franzose**
Franzose, blumenbachSammlung Nr. 382 (297); übermittelt durch Prof. v. Leveling

**Georgieni Caucasisch**
Schädel-Decaden III (1795; S. 39) Dissertation Blumenbach (deutsche Ausgabe, 1798, Verzeichnis, S. 2)

**Wolframmanul, Backenzahn**
1797, Handbuch der Naturgeschichte (6. ed., 1799: 121f.)

**Dazl-Porphyr Porfadro rosso aniko**

**Vulkanischer Sand aus der Lava des Tanaga-Vulkan**
Fundort: Tanaga Island, Androscoggin Islands - Aleutian Islands - Aleutians West - Alaska - United States - North and Central America - World

---

Meteorit von Mauerkirchen, Oberösterreich (ed. König Ludwig I von Bayern (1796 - 1868))
Edition of noise
How to start?

- What is my research question?
- What kind of data are necessary to get an answer?
- What are the main data standards in my research fields?
- How do I have to model the data?
- Are data stable and persistent or only temporarily available?
- What kind of infrastructure do I need to publish my results and research primary research data?
Translate into the digital languages
Ein Menschenkind also, auch von welchem Volke unter der Sonne es wolle — verglichen mit dem von allemenschafterlichsten Affen, ist allen schon hervorragend, den körperlichen Charakter der Humanität zu bestimmen, und die Ekel zu erregen, welche die Natur zwischen menschlicher und tierischer Gestaltung befestigt hat. Und so finden zwey mächtig Art der Abbildung dieser Bedeutung sich so sehr von einander auszeichnenden Schablonen auch in dieser Sammlung eine passende Stelle.

Der von einem Neger würde zwar äußerlich auch zugleich gefühlt haben, mit einem Blick zu zeigen, wie schlechterdings blafs der Zige sind, wodurch er sich von den Köpfen
nummern geschaffen werden können.
Das Speichern in Xanadu, Nelson die Möglichkeit, eine virtuelle
Nummer eines Dokumentes zu
Ausgabe

Scans issues
Making texts machine readable

- **Manual transcription**
  - Humans type letter by letter (with no knowledge of the language), but by two independent typists (double keying)
  - Guidelines, transcription tools
  - Compare the versions and correct the errors
    - decide on the best transcription of these cases
    - reduces errors
    - easy to calculate a measure of transcription accuracy
      (=inter-annotator agreement: percentage of words/letters on which both transcribers agree)
      - 99.997% precision

- **Automatic methods**
  - Optical Character Recognition (OCR, today improved by machine learning!)
Text structures
Manual Transcription

- Old manuscripts are often not written in an entirely linear fashion (margins, crossed out sections etc.)
- What to do with non-ASCII, non-local, non-Unicode characters/symbols?
- How to avoid deliberate or non-deliberate corrections, i.e. how to ensure faithful transcriptions especially of historical material?
- Manual transcription (double keying) is extremely time consuming and hence very expensive
OCR Errors

have a negative effect on information retrieval and hinder further processing:

- often affect word segmentation (→ tokenisation)
- miss full stops (→ sentence splitting)
- often deviate more from correctly spelled word than typos do
  - letter-to-letter mappings are less likely to be one-to-one (e.g., `m' → `iii')
  - are more likely to affect several places in a word
  - can be systematic (→ spelling error correction methods often don't work well)
- Postprocessing may be necessary to remove artifacts
Legal aspects

Copyright
- stresses commercial aspects
- can be transferred and sold
- can be held by a company

Urheberrecht
- stresses the relation between a creator and his/her work (i.e. the creative process)
- typically cannot be transferred other than to heirs
- can only be held by "natürliche Person", not by company
Text layout

- Two ways
  - Direct typing
    Dies ist **fett**.
  - Digital Markup
    <p>Dies ist <b>fett.</b></p>

Tags
<div1 type="Werk">
    <head>Die Leiden des jungen Werthers</head>
    <div2 type=Teil><head>Erster Theil</head>
        <p> [...]</p>
        <div3 type="Brief"><head>am 4. May 1771.</head>
            <p>Wie froh bin ich, daß ich weg bin! Bester Freund, was ist das Herz des Menschen!
                [...]</p>
        </div3>
    </div2>
</div1>
XML – eXtensible Markup Language

- Metalanguage for markup languages to define data
- Defines the overall structure of text data by document type definitions
- Searching and POS is possible
- W3C Standard, rely like HTML on SGML (Standard Generalized Markup Language)
Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of resources and training events for learning TEI, information on projects using the TEI, a bibliography of TEI-related publications, and software developed for or adapted to the TEI.

The TEI Consortium is a nonprofit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors.

Want to become active in the TEI community? Become a TEI Member, join a special interest group, sign up for the TEI-L mailing list, and come to our annual conferences and members’ meetings.

Last recorded change to this page: 2013-03-12. For corrections or updates, contact web@tei-c.org
TEI

About 450 Elemente (Tags) defined and categorized in 3 classes:

- Core Tag Set
  - For every TEI document
- Base Tag Sets
  - Chose one base tag sets
- Additional Tag Sets
  - Use as many as you want

[https://de.dariah.eu/tei-tutorial](https://de.dariah.eu/tei-tutorial)
TEI – Core Tag Sets (example)

Foreign language
John eats a `<foreign lang="fr">croissant</foreign>` every morning.

Mentioned
`<mentioned lang="fr">Croissant</mentioned>` is difficult to pronounce with your mouth full.

Term
A `<term lang="fr">croissant</term>` is a crescent-shaped piece of light, buttery, pastry that is usually eaten for breakfast, especially in France.
Genius rührtst du dich nicht?
TEI – Base Tag Sets

- TEI.prose: markup for prose
- TEI.verse: for poetry
- TEI.drama: for drama
- TEI.spoken: for spoken language
- TEI.dictionaries: for dictionaries
- TEI.terminology: for terms
- TEI.mixed: for documents of different types and genres
- TEI.general: as TEI.mixed, but only one base tag set
Erwin, (er springt hervor)

Weh mir!

Ich bin's.
Specific DTD

- 17th century DTD: leaflet, Emblem books, calendar, prayerbooks, ...
  - Metadata (from catalogues for the header)
  - Genre specific elements (title, picture, subscription, according to 17th century classification)
  - Indexing (indexing, full texts, elements of picture according to Iconclass)

http://computerphilologie.tu-darmstadt.de/jg03/opitz.html
Metadata

- Standardised description of metadata for documents
- Dublin Core Metadata Initiative (DCMI) since 1994

„The Dublin Core Metadata Initiative, or ‘DCMI’, is an open organization engaged in the development of interoperable metadata standards that support a broad range of purposes and business models.”

http://dublincore.org
Metadata – Dublin Core

Core elements

- ID: ISBN/ISSN, URL/PURL, URN oder DOI, ggf. Bibliographic Citation
- Technical data: format, type, language
- Content: title, subject, coverage, description
- Persons / Rights: creator, publisher, contributor, rightsHolder, rights, provenance
- Networks: source, relation, audience, instructional method
- Lifecycle: date (ISO 8601)
Metadata

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights
Photo documentation of artifacts

1. Information
2. Medium, system of media
3. Visual phenomenon
4. Object recognition
5. Pre-iconographic description
6. Iconographic analysis
7. Iconologic interpretation

H. Klinke (2013)

http://arachne.dainst.org/entity/1115686/images
Poligons vs. Point Clouds

Triangulation is an old and simple approach, which divides a curved surface into triangles (Thales). This method is good for small/medium scale artefacts (e.g. statues). It permits to reach high accuracy (0.05 mm) and a very dense sampling, very rapidly (300 000 points / s).
Keywording / Ontologies

ICONCLASS (http://www.iconclass.nl)


Getty Vocabulary (http://www.getty.edu/research/tools)

Art and Architecture Thesaurus (AAT, http://www.getty.edu/research/tools/vocabularies/aat/)
Content Based Image Retrieval (CBIR)

Indexing / Search via Visual Content like colour, texture, shape etc.

http://tineye.com
using colour information: http://labs.tineye.com
http://www.incogna.com
http://fusion.pixolution.de

Google Images
(http://googleblog.blogspot.de/2009/10/similar-images-graduates-from-google.html)

Sketch-Based Shape Retrieval
http://gfx.cs.princeton.edu/proj/shape/
Editions and Corpora

Machine intensived cultural studies

- High resolution editions of the canon
- Make the difficult accessible parts of the cultural heritage available
- Integration into portals and corpora
Editions are silos