Introduction to Digital Humanities

Day 1

Sophie Robert-Hayek

University of Lorraine

August 23, 2024

Outline

About me

About me

- PhD in computer science / artificial intelligence and its applications to complex systems;

About me

- PhD in computer science / artificial intelligence and its applications to complex systems;
- Currently a postdoctoral researcher at the MSH in digital humanities;

About me

- PhD in computer science / artificial intelligence and its applications to complex systems;
- Currently a postdoctoral researcher at the MSH in digital humanities;
- Mostly working on:
  - Stylistic analysis of New Testament text/manuscripts;
  - The application of philogeny algorithms to build the genealogical trees of manuscripts.

What are digital humanities ?

Definitions

Definition

Digital Humanities

Digital Humanities are an interdisciplinary field that:

- combines traditional humanities research with **digital tools**;
- applies **computer science/applied mathematics methodologies** to provide insights into humanities questions.

Definition

Digital Humanities

Digital Humanities are an interdisciplinary field that:

- combines traditional humanities research with **digital tools**;
- applies **computer science/applied mathematics methodologies** to provide insights into humanities questions.

DH brings together experts from a wide range of disciplines:

- **humanities**
- **mathematicians**
- **computer scientists**

to try to provide new answers and new angles to existing problems.

What are digital humanities ?

Recent advances in computer science offer **unprecedented** opportunities to

- **Generate**

What are digital humanities ?

Recent advances in computer science offer **unprecedented** opportunities to

- **Generate**
- **Explore**

What are digital humanities ?

Recent advances in computer science offer **unprecedented** opportunities to

- **Generate**
- **Explore**
- **Interprete**

What are digital humanities ?

Recent advances in computer science offer **unprecedented** opportunities to

- **Generate**
- **Explore**
- **Interprete**
- **Engage**

with data.

The integration of computer science with humanities disciplines promises an in-depth transformation **in research, analysis, and understanding of existing data**.

What are digital humanities ?

The convergence of computer science and humanities can lead to:

What are digital humanities ?

The convergence of computer science and humanities can lead to:

- A **paradigm shift** (*computational turn*) in research methodologies and outcomes, both for the humanities researcher and the computer scientists;

What are digital humanities ?

The convergence of computer science and humanities can lead to:

- A **paradigm shift** (*computational turn*) in research methodologies and outcomes, both for the humanities researcher and the computer scientists;
- Mark a new era of **interdisciplinary collaboration** to tackle research questions from different angles.

What are digital humanities ?

The convergence of computer science and humanities can lead to:

- A **paradigm shift** (*computational turn*) in research methodologies and outcomes, both for the humanities researcher and the computer scientists;
- Mark a new era of **interdisciplinary collaboration** to tackle research questions from different angles.
- **Widen research perspectives** through Open data initiatives and collaborative platforms.

A brief history of DH

The first wave

First wave: Emergence of Digital Humanities:

- **1940s-1950s**: Early use of computers for text analysis and birth of quantitative linguistics (stylometry, lexicometry ...).
- Limited by computational power and resources.

The first wave

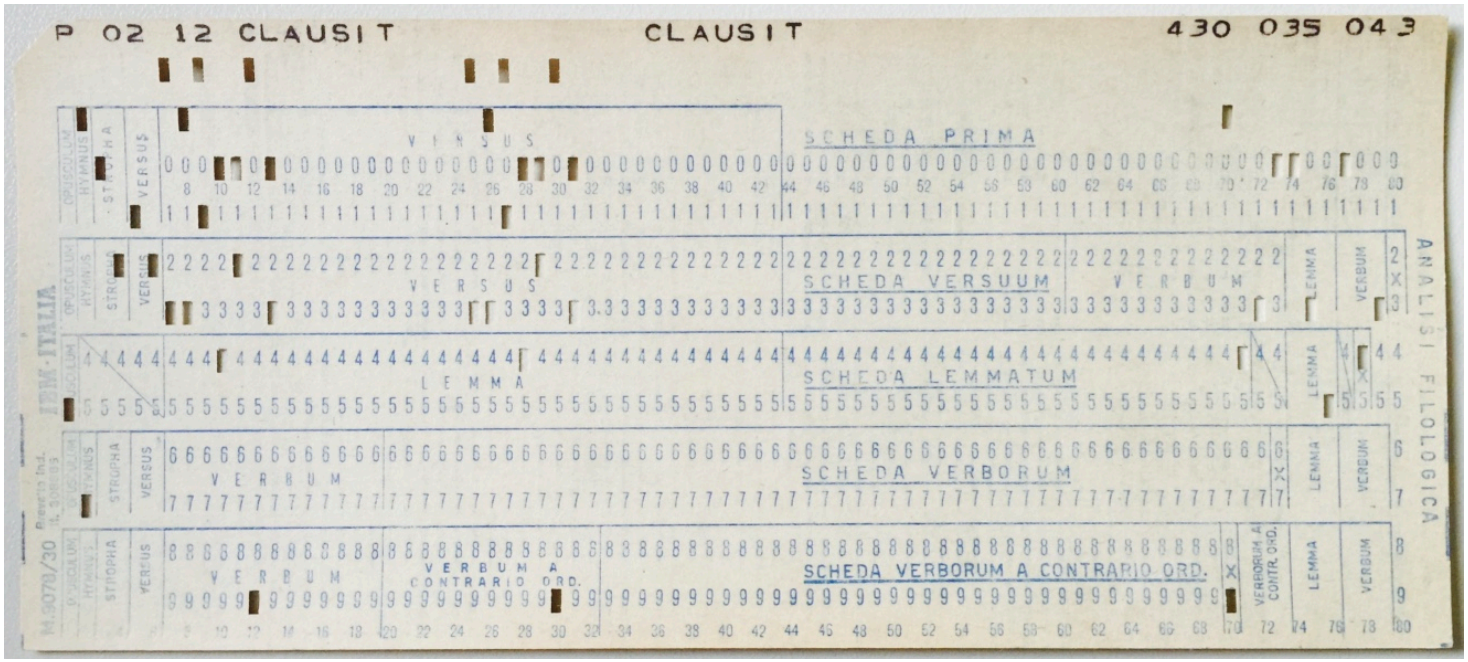
First wave: Emergence of Digital Humanities:

- **1940s-1950s**: Early use of computers for text analysis and birth of quantitative linguistics (stylometry, lexicometry ...).
- Limited by computational power and resources.

The first wave of digital humanities mostly dealt with **the application of quantitative models to textual data**.

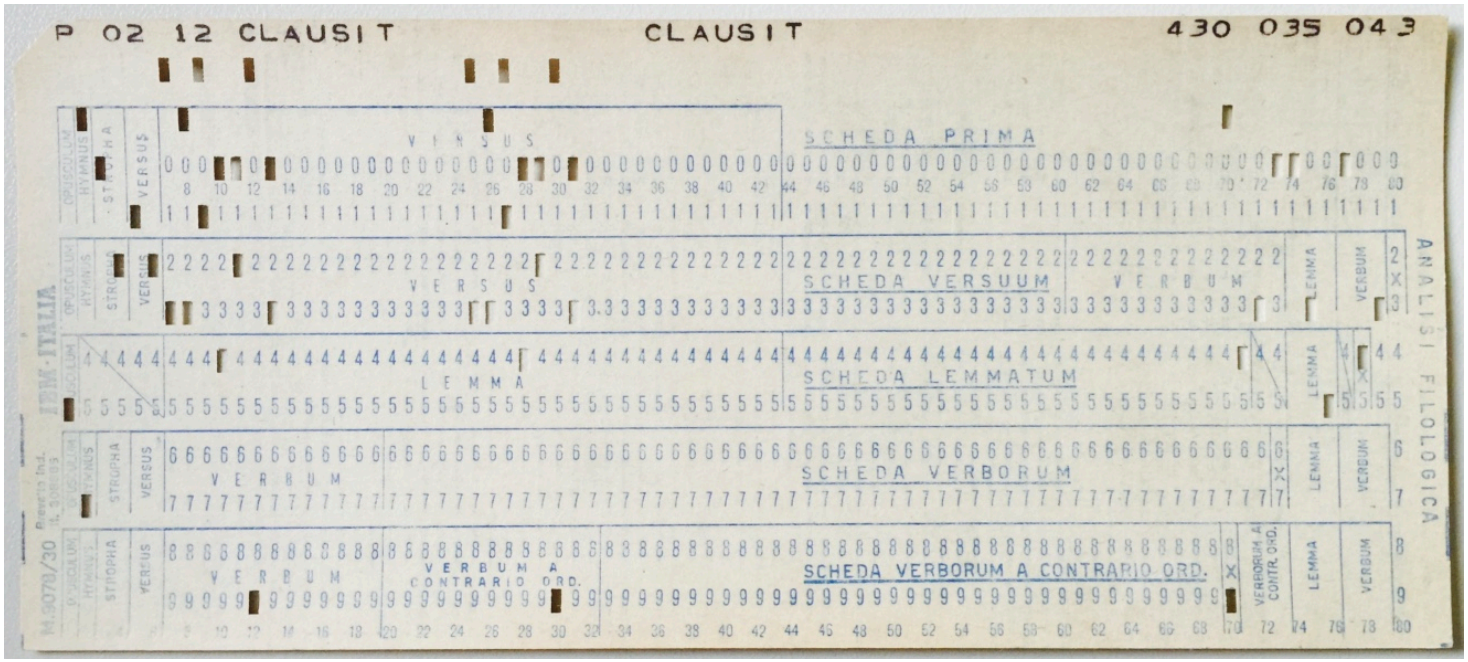
Punch card of Index Thomisticus

Roberto Busa's Index Thomisticus (1949-1970s) project aimed at building a concordance and a term frequency of words in the *Opera Omnia* of Thomas Aquinas.



Punch card of Index Thomisticus

Roberto Busa's Index Thomisticus (1949-1970s) project aimed at building a concordance and a term frequency of words in the *Opera Omnia* of Thomas Aquinas.



Each of us are going to do on Wednesday afternoon what took hundreds of people several years using punch cards ...

The second wave

The second wave (expression taken from Lou Bernard):

The second wave

The second wave (expression taken from Lou Bernard):

- **1960s-1970s**: Development of markup languages (SGML, TEI), for structured representation of texts.

The second wave

The second wave (expression taken from Lou Bernard):

- **1960s-1970s**: Development of markup languages (SGML, TEI), for structured representation of texts.
- Creation of specialized journals: *Computers and the Humanities* (1966), *Computer Applications and Quantitative Methods in Archaeology (CAA)* (1973), *Association for Literary and Linguistic Computing (ALLC)* (1978) ...

The second wave

The second wave (expression taken from Lou Bernard):

- **1960s-1970s**: Development of markup languages (SGML, TEI), for structured representation of texts.
- Creation of specialized journals: *Computers and the Humanities* (1966), *Computer Applications and Quantitative Methods in Archaeology (CAA)* (1973), *Association for Literary and Linguistic Computing (ALLC)* (1978) ...
- **1980s-1990s**: Creation of digital libraries and repositories, foundation for digital text analysis and archiving, such as The William Blake Archive in 1994.

The second wave

The second wave (expression taken from Lou Bernard):

- **1960s-1970s**: Development of markup languages (SGML, TEI), for structured representation of texts.
- Creation of specialized journals: *Computers and the Humanities* (1966), *Computer Applications and Quantitative Methods in Archaeology (CAA)* (1973), *Association for Literary and Linguistic Computing (ALLC)* (1978) ...
- **1980s-1990s**: Creation of digital libraries and repositories, foundation for digital text analysis and archiving, such as The William Blake Archive in 1994.

Going further than the application of quantitative models to linguistic data, the second wave **focused on the storage and the distribution of textual data**.

Third wave

Third wave: the appearance of the Web:

Third wave

Third wave: the appearance of the Web:

- Editions are now available online and easily shareable.

Third wave

Third wave: the appearance of the Web:

- Editions are now available online and easily shareable.
- Tools and software can now be transferred across teams.

Third wave

Third wave: the appearance of the Web:

- Editions are now available online and easily shareable.
- Tools and software can now be transferred across teams.

The emergence of the World Wide Web catalyzed the on-going momentum of DH, by adding the possibility to **communicate the results of digital scholarship**.

Current trends

The **data revolution** ($>$ 2000) and the **artificial intelligence revolution** ($>$ 2010) are **radically transforming digital humanities**:

Current trends

The **data revolution** ($>$ 2000) and the **artificial intelligence revolution** ($>$ 2010) are **radically transforming digital humanities**:

- Huge increase in computing speed;

Current trends

The **data revolution** ($>$ 2000) and the **artificial intelligence revolution** ($>$ 2010) are **radically transforming digital humanities**:

- Huge increase in computing speed;
- Improvement in storage capacity and technology;

Current trends

The **data revolution** (> 2000) and the **artificial intelligence revolution** (> 2010) are **radically transforming digital humanities**:

- Huge increase in computing speed;
- Improvement in storage capacity and technology;
- Improvement in data collection;

Current trends

The **data revolution** ($>$ 2000) and the **artificial intelligence revolution** ($>$ 2010) are **radically transforming digital humanities**:

- Huge increase in computing speed;
- Improvement in storage capacity and technology;
- Improvement in data collection;
- Huge improvements in Machine Learning models (birth of neural networks);

Current trends

The **data revolution** (> 2000) and the **artificial intelligence revolution** (> 2010) are **radically transforming digital humanities**:

- Huge increase in computing speed;
- Improvement in storage capacity and technology;
- Improvement in data collection;
- Huge improvements in Machine Learning models (birth of neural networks);
- Huge improvement in UI/UX designs and Web technology;

Current trends

The **data revolution** ($>$ 2000) and the **artificial intelligence revolution** ($>$ 2010) are **radically transforming digital humanities**:

- Huge increase in computing speed;
- Improvement in storage capacity and technology;
- Improvement in data collection;
- Huge improvements in Machine Learning models (birth of neural networks);
- Huge improvement in UI/UX designs and Web technology;
- Easier to deploy collaborative tools.

Digital humanities is one of the **fastest growing research trend** and promises to **bring new insights to a wide variety of problems**.

Taxonomy of digital humanities project

Digital humanities

Digital humanities projects can roughly be divided into three major categories :

Digital humanities

Digital humanities projects can roughly be divided into three major categories :

- **Application of Artificial Intelligence/Mathematics**: apply mathematical models to better understand humanities data;

Digital humanities

Digital humanities projects can roughly be divided into three major categories :

- **Application of Artificial Intelligence/Mathematics**: apply mathematical models to better understand humanities data;
- **Application of Data Engineering**:
  - structure data from physical/unstructured data;
  - define new data standards across the research community.

Digital humanities

Digital humanities projects can roughly be divided into three major categories :

- **Application of Artificial Intelligence/Mathematics**: apply mathematical models to better understand humanities data;
- **Application of Data Engineering**:
  - structure data from physical/unstructured data;
  - define new data standards across the research community.
- **Application of Software Engineering**:
  - develop software to facilitate the access and the manipulation of data;
  - design new way to engage with data to derive knowledge.

Digital humanities

During this week, we will learn how to:

- Manipulate datasets to **analyze word frequency data**;

Digital humanities

During this week, we will learn how to:

- Manipulate datasets to **analyze word frequency data**;
- **Compute** and **plot** statistical estimators on this data;

Digital humanities

During this week, we will learn how to:

- Manipulate datasets to **analyze word frequency data**;
- **Compute** and **plot** statistical estimators on this data;
- Manipulate textual data for **term analysis** across a wide variety of books;

Digital humanities

During this week, we will learn how to:

- Manipulate datasets to **analyze word frequency data**;
- **Compute** and **plot** statistical estimators on this data;
- Manipulate textual data for **term analysis** across a wide variety of books;
- Use Deep Learning to **numerize manuscripts**;

Digital humanities

During this week, we will learn how to:

- Manipulate datasets to **analyze word frequency data**;
- **Compute** and **plot** statistical estimators on this data;
- Manipulate textual data for **term analysis** across a wide variety of books;
- Use Deep Learning to **numerize manuscripts**;
- Perform **statistical testing on textual data** to test for authorship.

Digital humanities

During this week, we will learn how to:

- Manipulate datasets to **analyze word frequency data**;
- **Compute** and **plot** statistical estimators on this data;
- Manipulate textual data for **term analysis** across a wide variety of books;
- Use Deep Learning to **numerize manuscripts**;
- Perform **statistical testing on textual data** to test for authorship.

... so you can generalize these methods on your own datasets and your own projects !

During labs, do not hesitate to use your own data or ask questions regarding how you can transpose methods to your own projects !

Example of digital humanities project

OCR and paleography for manuscript analysis

Example of breakthrough: how many scribes hands in the Qumran Great Isaiah Scrolls?

Example of breakthrough: Analysis of Isaiah scrolls

- Groningen University showed that the handwriting of the scribe on the Isaiah scroll **changed enough to indicate another scribe**;
- **which was barely perceptible visually**;
- using Machine Learning algorithms;

Popović M, Dhali MA, Schomaker L (2021) Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). PLoS ONE 16(4)

OCR for numerization of manuscripts

OCR (Optical Character Recognition) enables the transformation of manuscripts, archives, and historical texts into **digital formats**.

OCR for numerization of manuscripts

OCR (Optical Character Recognition) enables the transformation of manuscripts, archives, and historical texts into **digital formats**.

Many projects aim at:

- **Improving existing tools**: Tesseract (which we will study ...), Octopus...

OCR for numerization of manuscripts

OCR (Optical Character Recognition) enables the transformation of manuscripts, archives, and historical texts into **digital formats**.

Many projects aim at:

- **Improving existing tools**: Tesseract (which we will study ...), Octopus...
- **Digitizing documents**: MiDRASH @ EPHE ...

OCR for numerization of manuscripts

OCR (Optical Character Recognition) enables the transformation of manuscripts, archives, and historical texts into **digital formats**.

Many projects aim at:

- **Improving existing tools**: Tesseract (which we will study ...), Octopus...
- **Digitizing documents**: MiDRASH @ EPHE ...
- **Providing easier standards and tools to improve the efficiency of OCR**: Alto XML standard `https://www.loc.gov/standards/alto/`, Transkriptus, e-Scriptorium `https://msia.escriptorium.fr/` ...

NLP/Text Mining projects

NLP projects

NLP (Natural Language Processing) projects can be divided into:

NLP projects

NLP (Natural Language Processing) projects can be divided into:

- Tools, software and resources to facilitate the analysis of textual data:

NLP projects

NLP (Natural Language Processing) projects can be divided into:

- Tools, software and resources to facilitate the analysis of textual data:
  - **Collation tools**: the collatex project;
  - **Text analysis tools for ancient languages**: Dikta project;
  - **Various online dictionaries** of common and rare languages.
  - **Tagged datasets** made available.

NLP projects

NLP (Natural Language Processing) projects can be divided into:

- Tools, software and resources to facilitate the analysis of textual data:
  - **Collation tools**: the collatex project;
  - **Text analysis tools for ancient languages**: Dikta project;
  - **Various online dictionaries** of common and rare languages.
  - **Tagged datasets** made available.
- Applications of computer science models to **better understand the texts**:
  - Analysis of vocabulary;
  - Analysis of stylometric features;
  - Analysis of relationships between texts;
  - ...

Example of NLP project: Did Molière write his own play?

Example of NLP project: Did Molière write his own play?

Close **stylistic distance** between Molière's and Corneille's plays has led to much debate regarding the authorship of Molière's play.

Example of NLP project: Did Molière write his own play?

Close **stylistic distance** between Molière's and Corneille's plays has led to much debate regarding the authorship of Molière's play.

- *Si deux et deux sont quatre, Molière n'a pas écrit Dom Juan*, Dominique Labbé, 2009.
- Florian Cafiero et Jean-Baptiste Camps, « Why Molière most likely did write his plays », Science Advances, vol. 5, no 11, 1er novembre 2019

Example of NLP project: How many textual sources in Genesis and Exodus?

Source detection in ancient texts

Many ancient texts are composed of several sources, and stylometry has hinted towards:

- **Several sources in Genesis and Exodus** (*A Statistical Exploration of Text Partition Into Constituents: The Case of the Priestly Source in the Books of Genesis and Exodus*, Gideon Yoffe and Axel Bühler and Nachum Dershowitz and Israel Finkelstein and Eli Piasetzky and Thomas Römer and Barak Sober, 2023.)

Source detection in ancient texts

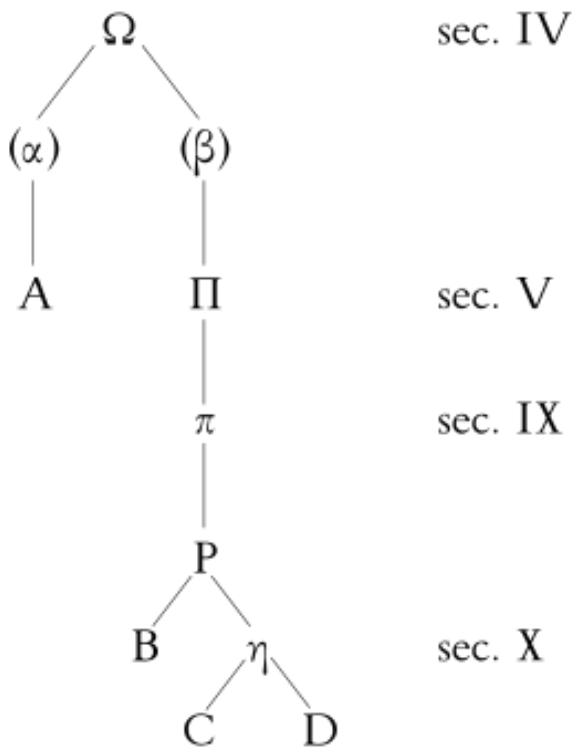
Many ancient texts are composed of several sources, and stylometry has hinted towards:

- **Several sources in Genesis and Exodus** (*A Statistical Exploration of Text Partition Into Constituents: The Case of the Priestly Source in the Books of Genesis and Exodus*, Gideon Yoffe and Axel Bühler and Nachum Dershowitz and Israel Finkelstein and Eli Piasetzky and Thomas Römer and Barak Sober, 2023.)
- **Several sources in the synoptic Gospels** (*Unraveling the Synoptic puzzle: stylometric insights into Luke's potential use of Matthew*, Sophie Robert-Hayek, Jacques Istas, Frédérique Rey, 2023.)

Stemmatology

Stemmatology

Stemmatology consists in finding the relationship between manuscripts to organize them as a genealogical tree.



Stemmatology

Traditionally, these stemmata are built manually (or semi-manually) by **a careful examination of variants**.

Stemmatology

Traditionally, these stemmata are built manually (or semi-manually) by **a careful examination of variants**.

I am a post-doctoral researcher on the SHERBET project, aiming at using **philogeny based automatic methods** to construct stemmatology trees of the Hebrew manuscripts.

Data initiatives

Archiving initiatives

- Gallica @ the BNF.
- Project Gutenberg.
- Brown Corpus.
- ...

Collaborative research projects

Example of virtual tools enabling crowd sourcing for manuscript transcription:

- New Testament Virtual Manuscript Room, @ the Institut für neutestamentliche Textforschung
- Scripta Qumranica Electronica, @ the Göttingen and Haifa Universities

Bibliography and questions

Bibliography

- *A Companion to Digital Humanities*, Susan Schreibman, Ray Siemens and John Unsworth, *Wiley–Blackwell*, 2004.
- *A New Companion to Digital Humanities*, Susan Schreibman, Ray Siemens and John Unsworth, *Wiley–Blackwell*, 2016.

Questions

Question ?

Questions

Question ?

Let's dive in after a short break !