INTRODUCTION TO INFORMATION VISUALISATION TECHNIQUES FOR HUMANISTIC SCHOLARSHIP:
A SERIES

⚲

This introduction was developed as part of a series on information visualisation for the humanities developed for the KU Leuven Faculty of Arts. The series was authored by Houda Lamqaddam, under the supervision of Prof. dr. Margherita Fantoli.

Find more on the work done in DH by the KU Leuven Faculty of Arts on the following page:
https://www.arts.kuleuven.be/digitalhumanities/english.

KU LEUVEN

Chapter 2
Getting to data

Starting from your own research material, one of the first steps you would have to take is to capture it as 'data'. This can be done by digitizing, extracting, or perhaps even counting elements in your material to save them in a digital format that you will be able to computationally process. There are a few possible methods to do that, and you can decide which to go for depending on the type of material you have. Typical techniques include:

- Optical Character Recognition (OCR): Use OCR software to scan and convert text from images or PDFs into editable text.
- Data Entry: Manually enter data from archival material into a digital format, such as a spreadsheet or database.
- Audio/Video Transcription: Transcribe audio or video recordings into written text.
- Metadata Extraction: Extract metadata, such as dates and keywords, from digital files.
- Digitization: Use a scanner or other equipment to convert physical archival materials into digital files.

This step is a critical one, as some of these methods may introduce errors (e.g. OCR), or might be very time consuming (e.g. audio transcriptions). The quality of your data – and therefore of your analysis – depends then on the adequacy and the quality of the data capture process.

There are other ways to 'make' data, and multiple steps to these procedures. See the VU's Digital Humanities Workbench for a detailed look into how to go 'From source to data'

Types of data

At this stage, it becomes possible to start thinking in terms of 'data'. The transformation that your material has gone through in the previous step allows it to be computationally *malleable.* But before we can do anything with it, take a moment to envision which type of data you are – or will be – working with. Does your material look like pictures? Maps? Text? Is there accompanying categories that would support in your analysis? What do they consist of? Have a look at the following data types, and see which description fit your project best.

- Numerical data: data that can be quantified and measured, such as a painting's height, weight, or a room temperature.
- Categorical data: data that can be placed into categories, such as a book's genre, or an objects' material

- Ordinal data: data that can be ordered or ranked, such as levels of education.
- Binary data: data that can only have two possible values, such as yes or no, true or false.
- Text data: data that is composed of words or characters, such as novels or online social media posts.
- Image data: data that is composed of visual information, such as photographs or sketches.
- Audio data: data that is composed of sound, such as music or speech.
- Time-series data: data that is collected over time, such as word usage.
- Spatial data: data that has a geographic or spatial component, such as GPS coordinates or satellite imagery.
- Relational data: data that consists of elements or people and relationships between them, such as correspondence networks.

◎ *For an in-depth exploration of 'data' in the humanities, read the recent article '[What do we mean by "data"? A proposed classification of data types in the arts and humanities](#)' by Gualandi et al.*

Again, let's take a moment to see how this applies to your own research material. What are you starting with? Which method would be appropriate to digitize or structure it? And what type of output (data) does it produce? Here are a few examples of data and methods that could be used for different types of research material:
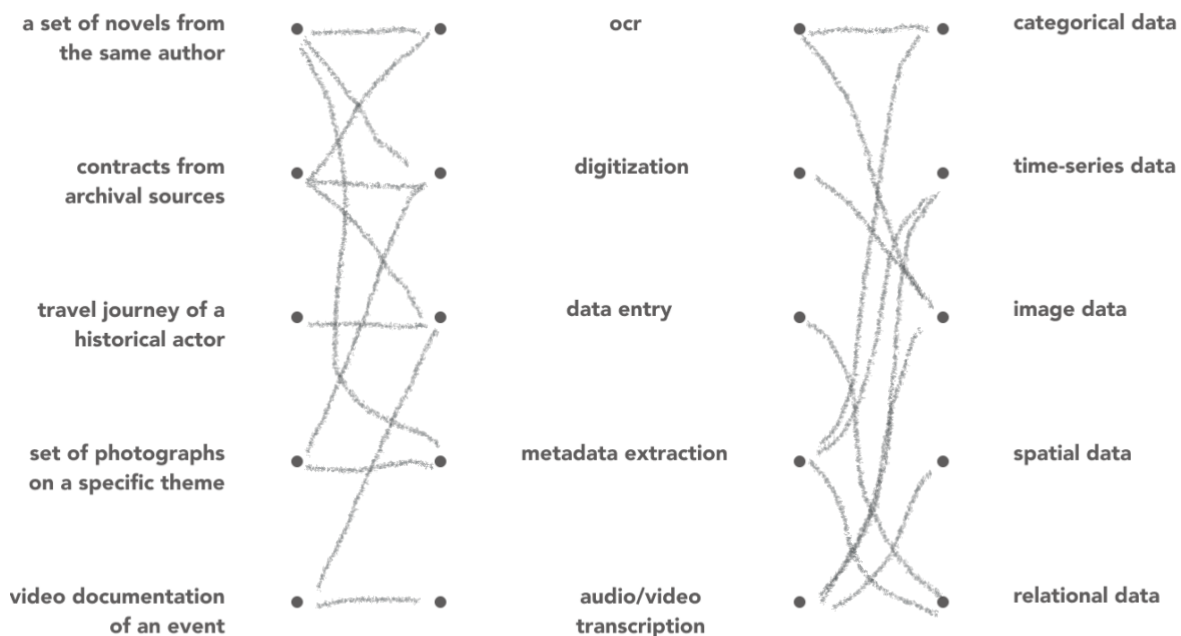
We have data. Now what?

The good part is that once we have structured data (i.e. data that has been organized to fit into one of the previously defined types). We can use some of the infovis guidelines that help us identify which type of representation would work best with our type of data. In the coming section, we will see how to do that practically.