Report 1e CODE-Congres: *Data, tools en output*

Date: 21 maart 2025
Place: Vrij Universiteit Amsterdam
Report: Dr. D.A. Sirks; Dr. M.A.J. Vermeer en drs. R. Haverman

1. Dr. Erika Kuijpers: The Chronicle Corpus

Dr. Erika Kuijpers (VU Amsterdam) presented the recently completed project 'Chronicling Novelty. New Knowledge in the Netherlands 1500-1850', which she led together with Judith Pollman (Leiden University) from 2018 to 2023, funded by an NWO Free Competition grant. The project involved two PhD candidates, Alie Lassche and Theo Dekker, who both defended their dissertations in 2024. The project is based on the idea of the revolution in information provision and science that took place in the early modern period. Kuijpers and Pollman proposed the question how new knowledge spread to and was received by the general population. To answer this question, they compiled a corpus of chronicles. On one hand, chronicles are a specific historiographical genre; on the other, they were the ideal place for recording and keeping observations from their own time and environment. Many early modern chronicles thus also served a dual function as knowledge collections. In total, the project collected no fewer than 204 chronicles in 308 volumes from the contemporary Netherlands and Northern Belgium. All chronicles were selected that written on the scribes' own initiative, not as commission by royal courts or monasteries,. These manuscripts were scanned via Vele Handen and DBNL, and made accessible through Transkribus. Volunteers were used to train models. For Lassche's doctoral research, the volunteers additionally annotated text passages containing observations at a detailed level. This proved to be a very time-consuming task: within the project's duration, about two-thirds of the text was ultimately reviewed and annotated. The researchers created a dataset of these chronicles. They were able to make some observations, such as the significant increase in historiographical activity during times of crisis. The material will be provided in the future to the institutions that preserve the manuscripts; the metadata, documentation, and datasets have been placed online.

2. Prof. Dr. Matthias van Rossum, *Colonial Archives, Lawsuits, and Social History*

Prof. Dr. Matthias van Rossum (IISG & RU) gave a lecture titled 'Colonial Archives, Lawsuits, and Social History – Some Possibilities of Indexes, Datasets, and Digital Infrastructure.' For example, for the GLOBALISE project, he uses the criminal case files of the VOC to study the social history of underrepresented groups on a global scale. These files had already been scanned and made available by the National Archives, but the accessibility was no further than inventory level. By indexing based on, for example, personal names, charges, and locations, the archives can be made digitally searchable. In a second cluster of data projects, Van Rossum attempts to uncover knowledge regarding the VOC's slave trade in Asia. For example, the individual legal documents found in the case files (such as transport deeds, or proofs of ownership transfer) can provide insight into the larger patterns of the Asian slave trade: who is sold to whom, and how old are these people? In this way, it is even possible to follow individuals over time. The goal of the projects is not only to create an extensive data collection but also to set up a digital infrastructure in which the case files are searchable, supplemented with semantic language technology and historical reference data.

3. Prof. Dr. Stephan Dusil, *The Consilia of the Tübingen Law Faculty*

Prof. Dr. Stephan Dusil reported on his research project on the *consilia* of professors at the law faculty of the University of Tübingen. Until the year 1879, these professors frequently wrote legal opinions at the request of individuals or courts. The oldest advices are preserved as expeditions; from the seventeenth century onwards, the faculty itself kept registers with copies. Additionally, some have been published in print. The thousands of advices cover all fields of law, making it a very valuable serial source for a multitude of disciplines, such as reception history, regional history, social history, and philology. Dusil took as a case study the register covering the summer deanery (*Sommerdekanat*) of 1808. He transcribed it with a group of volunteers and incorporated it into an extensive XML structure. The current stage of the project involves tagging various types of entities, such as personal and place names, and producing summaries with the most essential information. For practical reasons, Dusil has currently chosen to focus on publishing the data without first correcting it. With this 'dirty' data, the source would at least be opened for research. However, this choice has both advantages and disadvantages. The use of Large Language Models, in this case, ChatGPT, has led to somewhat successful results, but not entirely without problems. Dusil is therefore currently considering the best choices for the next stages of the project.

4. Dr. Rik Hoekstra, *The Web Application Goetgevonden*

Dr. Rik Hoekstra's contribution concerns the Goetgevonden project, whose digital environment was opened to the public in 2024. In this project, the resolutions of the Estates General were digitized. This archive is a key source for political and other history and has been preserved in serial form for nearly two and a half centuries (1576-1796). Although the oldest years were published in the Grote Reeks of the RGP last century, the corpus is so large that a manual approach is unfeasible: about 650 volumes with a total of approximately 1 million resolutions. The advantage of the serial nature of the archive is that it consists almost entirely of texts with a very clear structure, making partial automation of the process a possibility. For this project, the traditional structure of the archive had to be broken up. Instead of a physical structure dictated by the material form of the registers, a logical structure was needed. This meant that no longer was the page the base unit, but the meeting or individual resolution. In preparation, the National Archives scanned all the volumes. They were then made accessible. Step 1 involved transcribing the scans. This was first done with the help of volunteers via Transkribus, and later through recognition by Loghi. The second step involved segmenting the texts based on meeting days and individual resolutions, which could be automated by recognizing date formulations and formulaic language, with or without fuzzy searching to circumvent spelling variants and text recognition errors. The third step involved recognizing and indexing all entities, such as personal names, place names, and titles. With these quantitative, serial data, the researchers were well able to answer more qualitative questions. For example, the data were suitable for identifying changes in the handling of petitions and in individuals' access to the Estates General, identifying the influence of provinces on decision-making, or establishing a relationship between language standardization and a growing bureaucracy of the Estates.

5. Prof. Dr. Hylkje de Jong, *The HUF-project*

Prof. Dr. Hylkje de Jong (VU Amsterdam) concluded with a lecture titled 'The HUF Project.' This project encompasses a large corpus consisting of civil case files from the Court of Holland, the Court of Utrecht, and the Court of Friesland between 1700 and 1811, and involves a collaboration with the National Archives, the Utrecht Archives, and Tresoar. A pilot project has already started with civil cases that took place between 1716 and 1730 for the Court of Friesland. The chosen period is no coincidence: in 1723, the revised Provincial Ordinance of Friesland was proclaimed, and the selected cases thus provide insight into the application of the law in the period before and after the implementation of this ordinance. Although the initial plan was to process the civil case files, the technology proved insufficiently compatible. Due to the many annotations in the margins, automatic transcription by a computer model resulted in too many errors. For this reason, the decision was made to use the sentence books

instead, which are more suitable for training a computer model. The pilot project serves two purposes. Firstly, at the macro level, the sentences are automatically transcribed with a trained computer model and indexed by, for example, personal names and legal sources. This lays the foundation for a digital infrastructure. Secondly, at the micro level, the sentences are studied from certain legal themes, such as the offense of insult (*iniuria*). What does the future look like? Ideally, the HUF project will make the entire corpus accessible in such a way that it is searchable and enriched with links to similar cases from the corpus, as well as online available legal literature and primary source material.

Discussion: Four specialists in conversation

Following the morning lectures, the afternoon continued with a panel discussion. Four experts, moderated by Professor **Gijs van Dijck**, engaged in conversation. Whereas the earlier speakers had focused on data and output in their digital historical research, the panel shifted attention to the underlying technology.

The panelists were researcher **Katrien Depuydt** (Institute for the Dutch Language), Transkribus community director **Annemieke Romein**, KNAW software engineer **Rutger van Koert**, and project leader **Edwin Klijn** (both active for the CABR – Central Archive for Special Criminal Jurisdiction). In short, the panel consisted primarily of practitioners who shared their hands-on experience with technical tools.

Klijn opened the discussion. When asked by Van Dijck, "What are the possibilities of data and technology?", he responded that data serves as 'essential resources' for solid digital research. According to him, a good, standardized data structure is essential. Depuydt and Van Koert echoed this view. "Technology also allows us to uncover things that used to remain hidden," Van Koert added.

Connecting through technology

Van Koert, who previously worked as a *techie* in the commercial sector, is now involved in developing software to analyze historical handwritten documents (such as Loghi). Romein, too, is active in this field as a Transkribus-representative. "Honestly, I'm a bit lazy," she joked. "I don't feel like slogging through texts manually. Any tool that helps me transcribe is more than welcome."

Like Loghi, Transkribus uses *Handwritten Text Recognition* (HTR) software to decipher old documents. Romein shared that the newest model – 'Dutch Dean' – has been trained on no less than 25 million words. Still, she emphasized that understanding other tools and their applications remains crucial. "Technology connects. It's great that this conference brings together ideas and tools from various disciplines."

"So how do you actually get started with these tech tools?" Van Dijck then asked. "When uploading your material – such as scanned documents – it's important to understand what you're working with and to think about what you want to get out of

it," said Romein. "That allows you to identify different handwriting styles and structure the layout for optimal text recognition."

From content to technology

Klijn also stressed the importance of keeping the end goal in sight when making technical decisions. "The CABR is a massive judicial archive about individuals investigated for collaboration during World War II. We are developing a website to make this archive accessible to a broad audience. It is crucial that our technical choices are based on the question: what functionality should we provide to users?"

According to Klijn, the primary users are often relatives of the people involved. "They need clarity amid the complexity of digitized documents. Search terms should lead directly to relevant records that contain the results they are looking for. Moreover, we've reconstructed the legal proceedings and taught the computer to link key documents – such as decisions and verdicts – to those processes."

Van Koert was closely involved in the development of the CABR website, which took about a year and a half to complete. "We opted for automatic document type recognition using *text similarity search*. That allowed us to group similar documents – such as summonses, which follow a fixed textual format – and subsequently classify them."

KISS

Katrien Depuydt also adheres to the principle of 'look before you leap' in her work. At the Institute for the Dutch Language, she is responsible for building and publishing digital corpora and (historical) lexica. "You need to design your data infrastructure in such a way that you can build upon it later," she stated. "Make your datasets flexible and future-proof so that data modeling remains possible down the line."

When Van Dijck asked the panel members to share potential pitfalls, Depuydt followed up immediately. "People often want quick results, but it's important not to jump in too soon. First ask: what do I actually want to do with my data? If you skip that step, you'll inevitably face lots of post-processing later."

Romein concluded by reaffirming the importance of careful planning. "Keep it simple and straightforward," she said in closing. "Know your users, talk to them, and test continuously. Does the output still meet your goals? An 80-year-old who wants to digitize a diary has very different needs than a professional archive."