Original Author Manuscript of

Rik Hoekstra, Marijn Koolen
"Data scopes for digital history research"
Historical Methods A Journal of Quantitative and Interdisciplinary History 2018 (2): 1-16
DOI: 10.1080/01615440.2018.1484676

Data scopes for digital history research

Rik Hoekstra (rik.hoekstra@huygens.knaw.nl)
Marijn Koolen (marijn.koolen@huygens.knaw.nl)
Huygens ING - KNAW Amsterdam

Introduction

Researchers need instruments to collect and study their data. Astronomers use telescopes, biologists use microscopes. With the coming of big data, a number of researchers decided to use the term macroscope to qualify the instrument needed for studying data clusters, large amounts of data from one or more datasets. This macroscopic approach is very different from traditional historical research practice on which scholarly argumentation and citation is based, but is still under-theorized, as Ted Underwood argues (Underwood 2014). Historians conceive their data as a reaction to their research questions. They also interact with their datasets and interactively enrich and enlarge them as research progresses. Each step of selection, enrichment, and classification represents a choice that is based on explorations and interpretations of the data. These interactions change the data and are essential in understanding any subsequent analysis, which makes them part of research methodology, but there is little consensus on how these steps can or should be performed. Moreover, they are rarely reported and discussed.

Historical research practice is shaped around discovery in physical, analogue sources, where, in the quest to find relevant sources, scholars encounter many related and unrelated documents that help them to understand those sources in their wider context (Hitchcock 2011; Gibbs and Owens 2013; Putnam 2016). Now that many historians increasingly use digital environments and data, research practices shift. Digital environments offer different ways to search and discover sources, that, as Lara Putnam warns us in "The Transnational and the Text-Searchable" (Putnam 2016), make it easy to find highly relevant documents, but often present search results on their own without their context, missing the built-in contextualization of discovery in physical archives. This is also reflected in referencing practices that are still based on hard-copy, even when argumentation is based on a digital search process. "To take a single example of this disconnect between research process and representation, many of us use and cite eighteenth and nineteenth-century newspapers as simple hard-copy references without mention of how we navigated to the specific article, page and issue. In doing so, we actively misrepresent the limitations within which we are working" (Hitchcock 2013, 12).

Several scholars, including Fred Gibbs and Trevor Owens (2013), Ted Underwood (2014), Graham, Milligan and Weingart (2015) and Jennifer Giuliano (2017) and the working group on Arguing with Digital History (2017) argue there is a need for more transparency when using digital methods for discovery and analysis, so the research community can discuss how these methods fit in current practice or how practice can or should adapt to incorporate the various digital methods. We argue that especially the transformative nature of digital data

interactions require transparency that goes beyond the requirements of analogue research to cite the sources used. Groth et al. (2012) point out that with explicit descriptions of data transformations, the users of the datasets can assess “the context in which the data was created, its quality and validity, and the appropriate conditions for use.”

Transparency of approach, whether digital or analogue, requires describing the activities that constitute that approach, and how they are combined in a process to generate new knowledge or ways of seeing. We argue that there is a relatively short list of activities that make up most digital research methods, and that to achieve transparency, it is important to understand the nature of these activities. The types of data transforming activities that should be brought to the surface in presenting historical research are: selection, modeling, normalization, classifying, and linking. Together, these interpretive activities allow researchers to shift between different perspectives on data. According to Benjamin Schmidt, the importance of understanding digital tools is not at the lowest level of algorithmic detail, but at the level of data transformations (Schmidt 2016).

This paper presents these interpretive activities, how they are connected, how they shape perspectives, and how they allow shifting between perspectives, through an overarching concept that we call *data scopes*. A data scope represents the process through which different views on research data are created that are relevant to a specific research question. By explicitly describing this process, researchers allow their peers recreate that perspective and to critically evaluate it.

Many different fields are developing best practices for computation transparency (Arguing with Digital History working group 2017, 19). Several digital historians deposit software and scripts used in their research on GitHub for others to scrutinize and reuse. While this is an important step, additional description and especially methodological argumentation is needed to arrive at a coherent and commonly understood set of methods for doing digital history. First, scripts in a GitHub repository only partially describe the process, lacking any manual interactions with graphical user interfaces. Moreover, the process is iterative and non-linear, with a lot of experimentation and backtracking as scholars deepen their understanding of the data, which is rarely documented but crucial to understand which views and interpretations were chosen and which were discarded or ignored. Second, scripts often lack reasons for steps described and discussion of their consequences, and can involve complex tools that may no longer be available, usable or documented in a few years time and can perform a large number of transformations based on a single instruction. Incorporating digital methods in historical research practice requires a description of the process that focuses on why and for what purpose particular steps are taken, and their consequences.

It is certainly not our intention to argue that historical research should be reduced to data scopes. There are many other valid ways to interact with historical sources and even single datasets, that may be used in addition to or in combination with the data scope methods we propose. But as Gibbs and Owens (2013) and Giuliano (2017) already argued, if historical researchers engage with large scale data, or clusters of datasets, they should account for

their methods and there is currently a lack of both methodological transparency and a coherent view of methods. We believe that data scopes can conceptually fill this gap.

In the emphasis on transparency and explicitness we propose to extend source criticism as historians have applied it to traditional sources to the realm of (big) data and digital tools (see e.g. Fickers 2012). Source criticism in the digital world in our view should comprise both a critical examination of the datasets themselves and of the transformations researchers use to adapt their data to answer their research questions. We introduce our concept of data scopes, to point out that creating data scopes is guided by research questions and that this is a constituent and integrated part of (historical) research. We elaborate data scopes as a coherent set of methodological principles that characterize the interaction between researchers and their data and the transformation of a cluster of data into a research instrument.

Many of the methodological issues we present tend be abstract. Therefore we have chosen to illustrate them as much as possible with a case study, that we introduce after this introduction.

Case study: Dutch-Australian emigrants data scope

The aim of this data scope was to make a multi-perspective view of Dutch migrants to Australia. The data scope set out with a set of 51,525 system cards, that were compiled by the Dutch consulates in Australia from 1945 to the 1990s. The consulates recorded migration units, usually a family, but by no means always. The records consisted of cards with data about the migrant units, such as birth date of the principal migrant, migrant unit composition and all sorts of events that happened to the migrants after their arrival in which the consulates played one role or another. With some 50.000 migrant units, the file covered about ninety percent of all Dutch-Australian emigrants from 1945-1990, estimated at 180-190.000 individuals.

The files were transferred to the Dutch National Archives, digitized and made provisionally accessible by a simple database with names and travel dates. Most of the information was not available in the database, but just on the cards. We drew a sample of one percent of the cards and found they contain all sorts of information about migrants, their interactions with the consulates and events that happened to them after they arrived in their country of destiny.

The cards could not be read by machine, because of the mixture of typescript and handwritten notes. We decided that we would use computer assisted methods to make them readable whenever possible, but that we would resort to manual research when necessary. However, it was clear that there many more datasets about these migrants were available. The Australian government kept records in their immigration files, that were accessible through the Australian National Archives Record Search, available on their website. And several Dutch archives held records about them, that were compiled during the selection procedure conducted before they were elected to participate in subsidized bilateral

programs between the Dutch and Australian governments. Moreover, the migrants themselves left many traces in numerous museum and archival collections in both the Netherlands and Australia and in personal collections, that we hope to start linking in a later phase.

The data scope got its initial shape with the decision about modeling. There many possibilities for the basic unit of the model, as there were so many aspects. For instance, we could have chosen the stories of a selection of the migrants as a basic unit, or we could have departed from the institutional perspective and see how policy decisions related to the data, or we could have conceptualized Dutch-Australian emigration as a number of overlapping networks. But we chose to take the database with emigrant cards as our starting point and take the events on them as the basis for the (partial) reconstruction of the life courses of these people. In this way, the life courses, consisting of events, became the backbone of our data scope. The advantage of this approach was that most other information about the migrants could be linked to it, as almost all sources would have information about events and nearly all emigrants mentioned in the other datasets were present in the cards. Moreover, it gave a lot of flexibility in filling in parts of the data set for our own project, but also for other users, such as members from the migrant community. Although we explain the concept of data scopes using this example, any of the alternative modeling choices would lead to a similar set of steps that can be described by the same set of concepts of data scopes and data transformation activities, but the order of steps and choices made would be different, leading to different data scopes and interpretations. Our case study is mostly about structured data, which allows us to explain the concept of data scopes and interpretive data transformation activities with a relatively simple example. However, data scopes are not confined to structured data and similar steps would be required in another research project for, for instance, data obtained by Optical Character Recognition (OCR), geo-locational data or any other source. While this would demonstrate additional complexities for normalization or linking, the methodological principles would not change.

A key device in the construction of a data scope are what we propose to call data axes, which are formed by the common elements in different data sets that link data snippets from different datasets in a structured way and give each other context and structure the data scope. The life courses and the life events were very suitable as a data axis for this data scope, along with the places (of origin and of destination), the migrant travel dates and the migrant units to establish relations for individual migrants (van Faassen, Hoekstra and Ensor 2015). Most of the datasets were available in a variety of digital heritage institutions in the Netherlands and Australia. It was not feasible or even desirable to collect them into one large composite dataset.

The data scope we started to build was modelled to contain life courses and events. They consist of smaller elements - actors (people and institutions), locations, dates, the event description proper and an event type. In this way events are complex objects that may be queried and analysed along different axes, which allows researchers to easily shift between different perspectives and focal points in a data scope. The event axis ties dispersed information from a number of originally distinct data sets. The extended information in those

datasets was linked to the data scope. An schematic impression of the way the data are linked into the data scope is shown in figure 3. Apart from the modeling discussed above, as simple as it is, the schema involves most of the other actions involved in creating a data scope. We link different resources that were compiled with different purposes and extend and contextualize the data in the data scope. In order to *link data* about the migration event, we have to *normalize data* about the migrants involved (names, birthdates), ships (spelling variation in names) and travel dates (departure, itinerary and arrival date). There are two types of *classification* involved in this case study: First, the migrant cards contain information about migrant units, and the Australian immigration files and the shipping lists about individual migrants so we decided to match individual migrants with a migrant unit. Second, the events are classified by type of event, for example employment, housing, finance and legal. These classifications require a lot of domain knowledge and coordination between the project participants. They both have a transformative effect on the data and change the context, in a way that would not (always) be obvious from the separate records, and that reduces data complexity.

The life courses approach had another transformative effect on the data. It placed the emigrant cards at the center of out data scope and not, for example, the immigration files from the Australian National Archives or the files from the community. All additional material that we would link now had to be modelled as or at least linked to events and a collection of life courses and not, for example, primarily as structural analysis of a group of people or as a number of stories. Even if these analytical approaches were not excluded, the choice of the modelling determined the basic perspective of the data scope. The modelling choices highlight certain aspects, and push others to the background. For instance, modelling events highlights the individual decisions and interactions between the consulates, the migrants and third parties like employers, but does not explicitly pay attention to the policy background in which the consulates operated. This may be added later, but as an additional layer that has to be modelled separately.

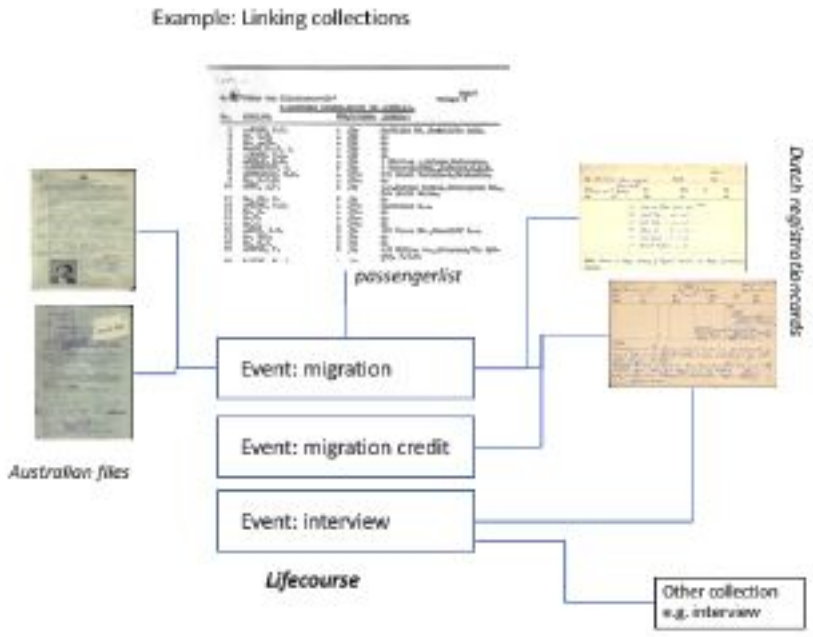

Figure 1. Example linking collections for life courses

Process of Digital Historical Research

The interaction between researchers and their data is of an iterative nature, in which exploration of sources and close reading lead to refined and additional research questions and data enrichment.

There are several different forms of interacting with data. For all but the most quantitative sources, there are forms of reading involved, but reading strategies require zooming in and out of details. In this way, researchers alternate between distant reading and close reading. Humanities studies have always known many forms of distant reading, using devices like (archival) inventories, bibliographies, thesauri, tables of contents or back of book indexes to get insight in and a grasp of a large body of materials (Blair, 2011). In addition, the digital age has brought many more sophisticated interaction tools with which to browse a body of material, such as fulltext search, tree views, faceted search and different types of structured access. Another way of interacting with the data are ways of aggregating data to observe larger patterns that cannot be observed without mapping and "reductive" transformations.

All these forms of distant reading serve to get to the relevant information efficiently. Usually, and especially with textual information, researchers need to read (or close read) that information afterwards. Close reading then gives them new ideas about and associations with other information or contexts. Or it prompts them to new research questions that may be answered with new cycles of searching, browsing and reading.

Apart from the well-understood cycle of searching, browsing and reading activities, there are various other forms of interacting with data sources, such as combining, aggregating or summarizing information and different forms of analysing, in which a researcher finds previously unobservable patterns and relationships in a resource. In their History Manifesto, Guldi and Armitage (2014) call for the identification of patterns and trends in large data sets, because they “open new possibilities for solving old questions and posing new ones” (88) They put forward that historians have several methodological and heuristic tools at their disposal, which gives them an advantage in handling big data sets: “noticing institutional bias in the data, thinking about where data come from, comparing data of different kinds, resisting the powerful pull of received mythology, and understanding that there are different kinds of causes.”(107-108). The order in which researchers interact with data are mostly unpredictable, which makes it all the more important to log the discovery and interpretation processes and make them traceable. Many research projects use the same types of interactions, but in each the choices and order of activities is unique, so without a detailed log, it's impossible to understand how interpretations are arrived at.

While digital resources make certain forms of searching information easier than ‘classical’ scholarly methodology, getting a grasp of the information contents of a large digital resource is notoriously hard. Of course, there is a difference here between classical and digital data sources and corpora. Paper data sources consist mostly of archival repositories, libraries, books and (printed or manuscript) texts and have a centuries old infrastructure of ordering and making information accessible, while making digital resources accessible has a tradition of at most forty or fifty years. Most of these methods still are derived from or analogous to the old scholarly methods of making data accessible, while there may be many more proper ways to regulate the interactions between researchers and their digital materials. With a few exceptions, however, these still need to be developed. Besser (2004) argues that this is the typical process of developing digital systems as complements or alternatives to analogue systems. The first versions of digital technologies are often a faithful copy of the analogue process, such as digital versions of the card-based catalogues in libraries. Gradually, the real affordances of the digital are discovered and implemented, thereby changing their roles and uses more fundamentally (Besser, 2004). At this point, research practice needs to adapt to these more fundamental changes.

It is a mistake to think that digital editions are in any way ‘the same’ as the analog sources they represent. For instance, photos and transcriptions of a text are both interpretations and different representations. Obviously, this is even more so if a source is turned into data, when many abstractions and normalizations are necessary to make data comparable. A few examples may clarify this:

- A photo representation of a source is a visual reproduction, that may, however, vary with many parameters. Lighting, reflections, the use of color and color settings, the camera, resolution and post processing like (lossy) compression all influence both (human) legibility and fitness for use of other tools
- There are many ways to transcribe a text. Traditions among branches of humanities differ a lot and range from an effort to transcribe ‘everything’ as faithfully to the original as possible to striving to produce a text legible for modern readers. If we include optical character or handwriting recognition and correction as a form of automatic transcription, it

becomes clear how many interpretative steps and (semi-)automatic transformations are actually required to obtain acceptable results.

- In digitizing a historical map, choices are made on which aspects drawn on the map to represent in the data and make operable. Turning a source into data requires modelling, which necessarily simplifies and distorts for the purpose of focusing on what is deemed most relevant. Which aspects are included in the model and which aspects are not is based on how the data is expected to be used.

In ‘classical’ source edition projects, that usually lasted several years, it was established that the only way to keep track of and account for the selection, elaboration and enrichment of sources, was to use formal criteria and record them (De Valk 1995). For the digital age, this has become even more important. Digital editing converts sources into data, often allowing near-instant re-organizing of the sources, which gives more flexibility but also more chances of losing track, not understanding what transformations have taken place and how to interpret the new organization. This is especially true for re-ordering based on non-formal metadata, such as fulltext search . It is easy to select all materials mentioning the same keywords, but it is not always clear whether they all refer to the same thing or are used in similar contexts.

The importance of interaction between researchers and digital data calls for a greater attention to the process and methodology of digital data interaction (Burke 2011, Huistra and Melink 2016). There seems to be no coherent methodology yet, but there are many best practices that have developed in decades (and sometimes centuries) of scholarship. In addition, computer science and tool development have also produced many useful tools, algorithms and practices that may save a lot of time and effort, but they also affect what is brought into focus and what is silently pushed to the background. As Putnam describes, keyword search often decontextualizes research and “deprives you of experiential awareness of just how rare mentions of your term were, of how other issues crowded your topic out in debates of the day” (Putnam 2016, 392).

Research focus

To answer research questions with historical data, researchers use either single datasets or clusters of related datasets. Usually these data clusters are seen as a source and a given, and they are not specific to any research question. But to create a historical narrative based on analysis of digital data, the scholar has to make choices in what data to select and use, how to organize, transform and analyze them, whereby interpretation takes places at every step. These choices affect how meaning is constructed, therefore are related to the research question that a scholar is trying to address. Answering research questions determines what is relevant in terms of the selection of data sets, what new links should be added to represent new knowledge, and in what way data points should be compared.

A high-level research question can be addressed in many different ways as it can often be interpreted in different ways. More specific and detailed research questions can be

answered more directly from the data, via specific lenses on relevant information axes. As explained above, data axes tie together common elements in different datasets so they contextualize each other and structure the data scope. Turning sources into data that can be used to answer research questions requires making them ready for querying and analysis. This is, as mentioned above, an iterative process consisting of many small steps that all transform and enrich the underlying data cluster. To illustrate this, in the events described on the migration cards from our case study, many consulate officials and other actors (employers, church ministers, social workers, etcetera) are mentioned, that are crucial for understanding the context of the migrants and the networks they were part of and that took care of them. Information about these actors can sometimes be gathered by combining scraps of information from the migration cards themselves, but often it is necessary to link the information to other sources, such as institutional information about the consulates from contemporary handbooks or from the institutional archival material kept at the Dutch National Archives. Church archives with both institutional information and personal archives from ministers or sources like personal letters contain dispersed information about the religious networks (that are different for each religion) and their backgrounds. Similar conditions exist for all types of individuals involved in the events. Of course, many different people were involved and only the key figures can be identified. This has important consequences for the data model that has to accommodate at least three different kinds of information: (1) more or less anonymous officials from many different backgrounds, as well as (2) serial information from institutional sources and (3) specific information the researcher gathers from dispersed sources. Putting them into one intricate model would greatly complicate the data scope but this problem can be solved by using an explicit model for core data about an individual, with links to unstructured (free text) data for other relevant information. This more specific information is often unique to a single source and does not have to be modelled explicitly as it will most likely not be used for structuring and querying data.
Transparent digital research includes argumentation for the steps taken and activities performed to arrive at a particular analysis and interpretation of the data, where the arguments are based on the research questions. Arguments for data selections can become complex when researchers include existing structured data sets, because they have to consider how their own research relates to and is affected by the process by which the structured data set was created. In the next section we discuss the type of steps and activities that are used to create a specific scope on historical data sets and how they affect interpretation. We can discern different types of activities, that together shape a data scope.

Selection

Selection is the activity humanities scholars have always documented most explicitly, also in the time before digital methods. In many ways, the selection of sources and of data has always been crucial and central to doing research. Selection can happen at any point in the process to add new data to a data scope, ranging from adding existing digital resources, but also digitizing collections that hitherto only exist in an analog form. In addition to the ‘classic’ forms of selection, of corpora, of documents and of parts of documents, the digital era has introduced some new, and sometimes subtle forms of selection that are still influential in shaping a data scope.

For example, searching an online resources, be it the ‘whole’ internet by internet search engines or a specific corpus with more specialized tools, yields results that are selected by a combination of the querying terms used and the particularities of the search engine. While the query can be made explicit, the way the search engine produces results for it is usually impenetrable, because its internal workings are not published or because the combination of indexing and retrieval and scoring algorithms makes it too complex. Moreover, searchers tend to focus on the first retrieved results, which, especially with many total results, introduces a bias to what the search engine considers most relevant (Joachims et al. 2007). If the search engine searches a corpus that is not under control of the researcher, changes in the corpus will make results unstable over time, especially in combination with changes in the search engine software. So, almost all web search queries from ten years ago that would be repeated today, would return vastly different selections.

As Laura Putnam remarks, digital search results are presented stripped from their context, making them hard to interpret (Putnam 2016). A more subtle selection problem arises when the corpus that is searched contains selections that are not clear for the researcher and (should) raise many questions. While this takes too many forms to describe here, some examples may illustrate these questions. For instance, in archives of national media outlets (newspapers, radio and television broadcasts) which specific outlets are included and which selection choices were made (Hitchcock 2013)? Are there any missing editions or broadcasts? Are there any that are too badly damaged to allow useful digitization? If the search engine offers full text search of OCR’ed materials, what is the variation in OCR quality across materials spanning decades or centuries? To what extent and in what way were OCR errors corrected? In composite corpora on the internet, how does the search engine deal with texts in different languages? Are non-western language materials included and if so, has there been an effort to reconcile e.g. author and place names? If the search engine index contains curated keywords (for example a library catalogue), to what part of the collection have they been they applied and how consistently? If an archive digitizes its collections, which parts are digitized? In what way? Is the whole collection searchable, or just parts of it? With all these questions the main problem is how scholars can tell what is selectable and understand how the provided methods of access shape their views.

Even many classic selection procedures would present a researcher with too many results to read or include into the research. An automated search often returns much even bigger selections. The ways of making digital search return smaller result sets, often differ substantially from analog research methods, as they include forms of extended querying: a researcher may choose to include only the results published most recently, or with a high user rating, or in a specific language, etcetera. All these practices are necessary and legitimate, but they should be accounted for, as they help to shape the scope of the data that is used to address research questions.

Modelling

Modelling according to Willard McCarty is "the heuristic process of constructing and manipulating models", where a model is "a representation of something for the purposes of study, or a design for realizing something new" (McCarty 2004, 255). Examples of modelling are identifying events and persons in migration cards to understand and analyse the interactions between migrants and their environment, or establishing relations between actors in early modern creative industries to investigate how these industries flourish in some places but not in others. Modelling is a core activity of any digital approach, as "models are fundamental to computing: to do anything useful at all a computer must have a model of something, real or imaginary, in software." (257) and to use a model demands "complete explicitness and absolute consistency" (255) To search through a digital image archive, a computer needs both a representation of the images (as sequences of bits on a disk and as sets of color values and coordinates for displaying on a screen) and of any metadata structured in a fixed set of fields to allow a computer to retrieve images in response to query. Higher level analyses, for instance of life courses that connect persons, locations and events across heterogeneous sets of digital sources--be they text, image, audiovisual or geographical data--require higher level models.

According to Flanders and Jannidis, data modelling is a critical tool and a central activity in Digital Humanities research (Flanders and Jannidis 2016). The increasing usage of digital tools shows that many scholars make use of explicit models, but often the role and impact of model construction go unreported in digital historical research. To create a data model that is relevant to a research question, and transform digital source materials to fit that model, requires many modeling assumptions and data wrangling steps. For instance, what happens with metadata records that are incomplete, how are uncertainties in names and dates dealt with, to what extent are names disambiguated and variants mapped, how are similar data dimensions (e.g. entities, events, topics) in data sets with different provenance harmonized? Treating these steps as mere data preparation removes an important part of the interpretive process from scholarly debate.

We argue that it is not enough to describe or publish the data models themselves, as the process of constructing them is important as well. Computational models, "however finely perfected, are better understood as temporary states in a process of coming to know rather than fixed structures of knowledge." Computers "are essentially modeling machines, not knowledge jukeboxes." (256) It is therefore not only important to describe the models that are used in digital history research, but also the modelling process and how it contributed to increasingly nuanced and insightful interpretations.

Explicit modelling of data is often done based on one or multiple schemas, with the goal of retrieving information from it. This determines the dimensions in the data that are relevant for the research, and requires abstracting the data; modelling should emerge from the research questions. The schema describes which elements in the source data are relevant and how they are related to each other and in this way it defines the dimensions that will be used as data axes in the data scope. This in turn requires intimate knowledge of and experience with

the sources. Without a solid understanding of the sources, it is hard to know what to model and to describe it in a schema.

Modelling is an iterative process, with many additional activities to transform data, e.g. selecting, normalising, linking and classifying, during which thinking errors are uncovered and subtleties in the data are found that suggest changes to the models. "In the initial stages of use, th[e] model would be almost certain to reveal trivial errors of omission and commission. Gradually, however, through perfective iteration trivial error is replaced by meaningful surprise." (256)

As researchers gradually develop their knowledge about a research topic and adjust their understanding of it, they need to adjust the schema. Because of this, only the parts of the data cluster that are needed for answering research questions should be subject to modelling. In many cases, researchers have the tendency to over-model their sources, trying to include all phenomena they encounter and add many dimensions to the model that add little to interpretation but make it more difficult to grasp and analyse. The results of this are a convoluted schema and data that are both difficult to comprehend and hard to query, which defies the objective of information retrieval. As resource data sets are often incomplete, it pays to keep models simple. As an example, we are building a ‘charterbank’ of all Dutch charters from the Ancien Régime to collect basic information about charters dispersed over many different archives in a single place and to enable researchers to do serial research on them, something that was previously impossible. The basis for this data scope are the charter descriptions in digitized archive inventories compiled by many different archivists over a very long period. The variation in these descriptions is enormous and in modelling we decided to use a very simple, restricted model, in which only the inventory number, the date, the inventory description and a link to the original url were compulsory. We do include information from other fields such as the regest (abstract) and a thumbnail of scans for searching if they are available. Making more than these aspects compulsory would exclude the majority of the charter descriptions, because usually available charter information is very summary. Bringing more into the model would either require all sorts of transformations to harmonize the differences or result in a resource that would be hard to query because of all the deviations and exceptions. But even more importantly, most charter descriptions from the various digital archives are very simple and extended information is only available in a small minority of cases. In this way, a model that would cover all sorts of exceptions and extended information would raise expectations about information accessibility for all charters in the charterbank, that cannot be met.

Modelling can be seen as the activity to create a frame of reference, such that heterogeneous data can be mapped or transformed to fit this model, whereby many assumptions are made, and create data axes on which resources can be compared.

Normalization

Normalization of historical data is the process of bringing surface forms expressed in data back to an underlying standard form. This is one of the older problems in digital history, that

was already present in the ‘classic’ use of digital methods in history, for the ‘cliometrics’ type of research. For instance, in the introduction to the volume about Dutch shipping on Elbing, Thomas Lindblad writes: “The structure of a database needs to adhere as closely as possible to the form in which the information is given in the source so as to make data entry and checking easier. It is advisable to keep modifications and calculations at the stage of data entry to an absolute minimum. The database thus becomes a near-replica of the source in machine-readable form. This approach, however, implies that a fair amount of preliminary processing has to be done before any statistical analysis can be undertaken, e.g. with respect to standardizing alphanumeric information, as well as linking individual entries and aggregations of numerical data. As a result, the original information is often used in a modified form. This, in turn, may be accomplished by either replacing the original data or by adding supplementary variables.” (Lindblad 1995, 394-395)

Computer science has provided us with many sophisticated methods that are not confined to a single database. Especially the use of linked data technology promotes the addition of normalized data as an extra layer in a data cluster. An example of the events in the Dutch-Australian emigration case study illustrates this. The emigration travel by ship was recorded in different sources, among which Dutch emigration cards, Australian immigration cards and shipping lists. In many cases the travel dates for what must have been a single travel differ slightly or even considerably between different sources for a variety of reasons, mostly related to human error either in the original documents or in manual transcription as part of the digitization process. Reducing variation requires several steps, first establishing a canonic list of ship travels, and then linking all emigrants to the appropriate journey. For a large part we can employ digital methods that may correct small variations in dates and triangulation of evidence, but if deviations are too large (for example a typing error in the year), human decision is required.

The use of data sources with different abstraction levels makes data clusters layered. Where sources introduce data with different origins to the cluster, adding abstraction layers adds additional interpretations to the cluster, that transform the view of the underlying sources. This should be transparent and explicitly considered in the source criticism.

Linking

Linking is establishing explicit connections between objects in the data sources that are not there from the beginning. This includes processes like identification, deduplication and normalization, all aimed at extending information about a single object or phenomenon (Klein 2017). Any form of hierarchization may also be performed by linking, for example to an authoritative source, but should be seen as a form of classification. Thus, taking a rather trivial example, one source may contain the birthdate of a person, while another contains data about occupations and a third has data about dates of death. Linking them to a person integrates these snippets of data without changing the original data of any of the sources. Another result of linking may be to connect different witnesses of an event from multiple sources, thus diminishing (human) errors or hiatuses that exist in a single resource. The aim is often to harmonize heterogeneous datasets to create a single integrated, consistent and unambiguous dataset (Ashkpour, Meroño-Peñuela and Mandemakers 2015). As historical

materials are rarely complete, consistent and unambiguous, harmonization in the context of historical research is only partly attainable and each process of harmonization should be tailored to a specific (set of) research question(s) (de Boer et al 2015).

Linking data from different sources and different data types has an impact on how different interpretations become less or more salient. As a case in point, in the archive of letters between William I, Prince of Orange and his correspondents, frequency-over-time analysis reveals a dip in the correspondence between William of Orange and his cousin Günther, Earl of Schwarzburg, around the year 1570 (Hoekstra 2007). Considering the data by themselves allows for many interpretations. Connecting these data to geographical information of William's whereabouts brings certain interpretations to the front, while pushing others to the background. In this case, the knowledge that William was staying with his cousin Günther at the time foregrounds a simple explanation of this dip.

Linking data in a data cluster is an interactive process, that is guided by the research focus. Usually, there is no blueprint for the eventual shape or content of a data scope. Similar to the other activities, linking data inside a cluster is an iterative and non-linear process involving different stages of discovery, normalization, data ordering, classification and extending the data cluster with datasets that provide additional information, a missing viewpoint or context.

Data sets can be linked according to different discipline-specific theories or frameworks and on many different axes, which can be specific and concrete (persons, organisations, roles, objects, locations, dates or events), or more implicit and intangible (themes, topics, genres, sentiments, political views). At the same time, possibilities for data linking across datasets depend on the schemas of the datasets as well as the way data have been entered into the structure of the schemas. While there are ways to link qualitative data or data with a lot of variation, different layers and chains of imperfect transformations make data clusters impenetrable for researchers. To return to our events example, if we have an extreme case in which we use (imperfectly) OCR-ed descriptions of events, and use Named Entity Recognition (NER) tools to mine these for names of institutions and persons, but do not have the opportunity to disambiguate and identify these, then linking these NER names to other resources results in an unreliable resource with many faulty and missing links.

Linked Data offers a convenient way to link varied data and datasets so they can be semantically queried. To make linked data work for effective querying and analysis, it is usually necessary to add several layers of structuring and equivalence. From this necessity have come efforts to conceptualize the structured parts of data linking as *scientific data lenses* (Brenninkmeijer et al. 2012). A scientific data lense is a view on a linked dataset that determines that certain sets of entities or concepts represent the same thing. An example is the linking and normalizing of occupational roles in historical census data (Ashkpour, Meroño-Peñuela and Mandemakers 2015). Occupations and their terminology change over time, but depending on the purpose, some occupational or terminological variations should be considered the same or not. Grouping different terminology representing the same occupation requires a research-specific lense that explicitly models which occupations are equivalent in the context of the research question (Zijdeman 2016).

Lensing allows multiple views for specific contexts of a researcher, a research question or project, without changing the underlying data. A scientific data lense is different from a data axis. A data axis represents a data type that occurs in multiple datasets and can serve as a dimension on which these datasets can be linked. A scientific data lense typically determines the equivalence of specific entities. With digital data, such lenses can be stored, shared and repeatedly used to provide different users with the same sets of views on the data. Ockeloen et al (2013) describe a methodology for keeping provenance.

To illustrate the concepts of data scope and scientific data lense and how they are related, consider an network analysis of the network around the Dutch-Australian emigrants. Among the people in the network are subsequent ministers from the Dutch Reformed Church, who were themselves both emigrants and part of their church in which they performed social and religious services for the migrants that belonged to it, but they were also part of the community. To study the role of the ministers, the researcher can define a scientific lense that includes an equivalence relation between the two minister, treating them as a single entity for analysing the relation between the Dutch Reformed Church and its minister in Australia. This equivalence relation is valid within the context of this specific research angle, but not necessarily in different contexts. The equivalence changes, and thereby the scientific lense, when the focus is shifted to the communication between the ministers and their families in which they are no longer representatives of the church but two family member among different families and migrants among other migrants. So the equivalence is specified to their role as representatives of a church, while in another context they primarily migrants. If these two analyses are both used to address a single research question regarding the Dutch-Australian migrants, the two scientific lenses are part of a single data scope.

Classification

Classification is the reduction of complexity by grouping (data) objects into predefined categories, or classes. It “serves two purposes, each important: by grouping together objects which share properties, it brings like objects together into a class; by separating objects with unlike properties into separate classes, it distinguishes between things which are different in ways relevant to the purpose of classification.” (Sperberg-McQueen 2004, 161) There are many different ways to accomplish this, using external classification schemes. This is always interpretative, whether existing classification schemes are used or new schemes are developed during analysis. Classifications act as interfaces that mediate between researcher and data, making “some kinds of information more accessible and some less.” (Erickson 2013, 140) Therefore, classes and properties used to classify should be relevant to the research question.

For example, the registration cards of post-World War II Dutch-Australian migrants record all interactions between the migration officers and migrants. They range from passport and visa issues, to intermediation for migrants who lost their jobs, fell ill or committed a crime. To get a grip on the events, we classify them into eight different categories (like legal issues, finance, employment, health, etc) which reduces complexity and makes it possible to count and compare. Such classifications allow both quantitative analysis (how often were migrants

confronted with legal issues, or how many of them) and qualitative (which legal issues came up and at which points in emigrants' life courses).

A typical problem with topical categories is that there are almost always ambiguous cases that either fit in multiple classes or in none of them or where it is hard to decide clearly whether they belong to a certain class or not. "Classification schemes necessarily involve some theory of the objects being classified, if only in asserting that the objects possess certain properties." (Sperberg-McQueen 2004, 162) There are many assumptions underlying classification, which, apart from theoretical assumptions, include the relevance of classes to both the dimensions of the research question and research data, the relevance of properties in defining class membership, and the assessment of whether an object possesses a property or not or to some extent that is enough to be considered a member of that class. Scholars should carefully consider these assumptions and report these considerations where they are non-trivial. "At the extreme, the assumptions underlying a classification scheme may become effectively invisible and thus no longer subject to challenge or rethinking; for purposes of scholarly work, such invisibility is dangerous and should be avoided" (Sperberg-McQueen 2004, 162).

Data axes with multiple facets of classification (e.g. the person axis can be organized on occupational role, age group, gender, place of residence, decade of birth, etc.) make it possible to reduce complexity and create a focus with specific views on the data using combinations facets. Faceted classifications allow a data-driven process of creating data scopes, by selecting data based on a combination of facets from multiple dimensions or taxonomies. Upon discovering a digital resource of interest, historians can use a combination of facets to prepare a scope on the data cluster to contextualize that resource along related data axes. For instance, zooming in on part of a data cluster based on certain historical periods and regions and a set of occupational roles of persons, provides a focus on the interactions of people in specific professional domains, such as female Dutch emigrants with an agricultural background migrating to Western Australia in the 1950s.

In historical research on large digital datasets modelling and classification are often grounded in what a historian discovers as being relevant dimensions in the data.[1] Ansley Erickson asks: "How can we organize information and keep it accessible in ways that will facilitate our ongoing thinking and writing, if we acknowledge changing focal points or areas of interest?" (Erickson 2013, 133) Over the course of a research project, the research questions and focus and the understanding of the materials may shift, requiring shifts in classification. Generic classes such as person, location, date and event are robust against shifts in research focus and data modelling, but may not be specific enough to be useful in bringing out relevant new meaning and interpretations. More specific classifications on the other hand are more susceptible to such shifts, and may become obsolete or require redefinition of classes and reclassification of data, which in turn would require further reflection on and discussion of the underlying assumptions.

[1] In this sense, a bottom-up construction of relevant dimensions or facets is related to bottom-up coding in grounded theory (Star 1988, Glaser and Strauss 1967)

Automatic classification and big data

The interpretive aspect of classification is important when considering the volume of data to be classified. Small datasets allow more detailed modelling and classification than very big datasets, mainly because of the amount of work involved. Christof Schöch points out an important distinction between smart data and big data, where smart data is "semi-structured or structured, clean and explicit, as well as relatively small in volume and of limited heterogeneity" and big data is "relatively unstructured, messy and implicit, relatively large in volume, and varied in form." (Schöch 2013) Automated classification, such as done by Named-Entity Recognition and statistical Topic Modeling tools for text, and face and object recognition tools for images and video, offers a way to add explicit structure to big volumes of data. The quality and usefulness of the classification is of course partly determined by the underlying quality of the data (bad OCR quality often leads to unusable classification results), but the main point is that these tools incorporate the same aspects of modelling, selection, normalization, linking and classification. Topic modelling is a tool described by Graham et al. (2015) as part of the historian's macroscope to automatically detect topics in a large collection of text documents, and classify words in these documents as belonging to one of these topics. This technique requires a definition of the elements in the data to identify (e.g. terms are single words or multiword phrases), selection (e.g. which documents to include, which terms can be considered stopwords given the nature of the data), normalization (deciding which punctuation is irrelevant and can be removed, and which variations of word forms to conflate to a single stem), linking (of occurrence of the same word across multiple texts) and classification (assigning each term occurrence to the most appropriate topic). Intimate knowledge of the digital sources helps scholars to make informed choices in configuring the tools and interpreting the outcomes, that is, in shaping the resulting data scope so that it offers useful perspectives for the research questions and goals.

Data scopes

Data collections and data clusters are mostly static concepts. While they may contain many different patterns and connections, most these remain latent until they are made explicit by some form of ordering, categorization, or analysis and interpretation. All these actions require some form of interpretation that is guided by the focus of the research. A data cluster is usually relevant for multiple research questions and different perspectives. Therefore, methods that support querying and analyzing a data cluster should cater for different views and different questions, even if this may require new iterations of enrichment and data linking. Data scopes transform the relevance of a generic data cluster towards specific questions. Moreover, they allow source criticism from different perspectives by zooming in and out along different data axes, to identify and call into question various artefacts of the data.

For research, a data cluster may be a starting point, but it is elaborated, transformed and enriched in the course of a research project. We do not pretend that there is a one-size-fits all solution or that this is even desirable. What we do propose is to treat collecting, comparing and elaborating data and data sets as a coherent and essential part of research and to create methods for systematically incorporating all these steps and provide the possibility to critically examine them. We would propose to conceptualize this whole process as data scopes, in extension of the metaphor of the macroscope as used by Graham et al. (2015).

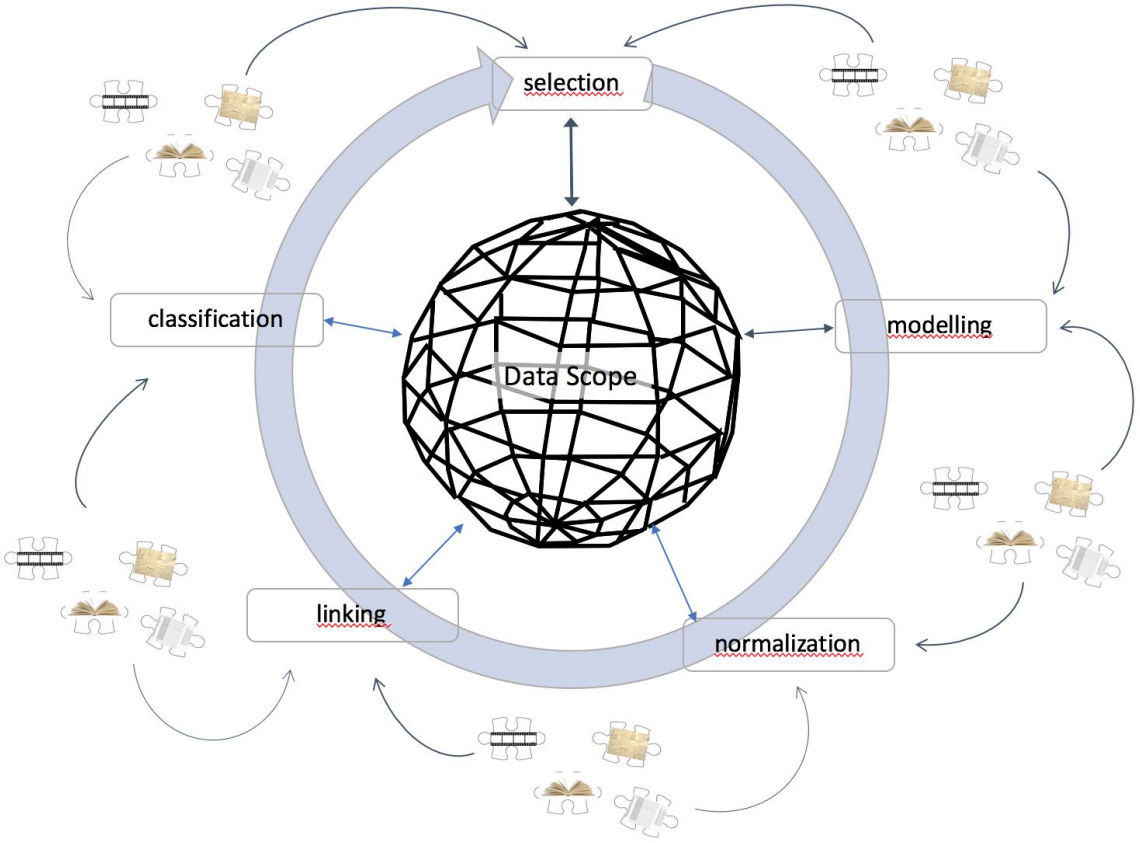

Figure 2. Schematic visualization of the iterative process of creating a data scope

The starting situation in creating a data scope is a researcher with a research theme or topic and a high-level research question, who approaches a data cluster as basic material for research. Transforming a data cluster into a data scope is a stepwise and iterative process, in which the data resource is enriched and transformed by the researcher's activities. For other researchers to assess a data scope, each of these steps should be documented as annotations of provenance on the data (de Boer et al. 2016), and described in addressing a research question. As Ben Fry explains about visualizing data, it requires a long chain of steps to arrive at a visual representation that provides useful insights, but it is crucial to be transparent about the process that led up to the visualization (Fry 2007), otherwise it is impossible for others to understand what they see. To refine a research question, researchers have to identify the dimension of the information in the (data) sources that they will use and modify for their research focus. This is a process of defining data axes in the data cluster. Productive data axes have the ability to link different data resources and to turn

data into information that may be queried and analysed and combined with other axes. The axes form the grid of information that a researcher employs to structure questions and to further explore the information. Usually, the resulting grid does not exhaustively cover the information in a resource, but expresses research perspectives.

Describing the reasoning behind data interactions and choices should be guided by the research questions. Owens (2014) argues: “What is the purpose of research questions in the humanities? I would posit that the purpose of them is to clarify what is in and out of scope in a project.” This scope includes the materials that are studied, the methods with which they are studied and theories and frameworks that structure and give meaning. We build on this argument and consider data scopes to be connected to the scope determined by the research question. They are interdependent. If the data scope is extended or shifted, this has an effect on the research question, and vice versa. Explicating data scopes is a way to check the research work against the research questions. Explication also allows more meaningful reflection on which data and interpretations are brought into focus and which are pushed to the background (Piersma et al. 2011).

The exact choices that the researcher makes, determine which sources (e.g. images, maps, texts, tables, audiovisual materials) end up in the selection, and therefore should be made transparent in research output to understand how the construction and use of a data scope led to the historical narrative. At the same time, the choices are hard to interpret without explanation, so publishing the choices made to create a data scope requires a motivation to understand a researcher's reasoning. Blanke and Hedges argue that linking data sets so that they appear as a single virtual data resource requires “significant understanding of the underlying semantics of the data” that are “for the most part left implicit” (Blanke and Hedges 2011, 656). Explicit data models make this understanding transparent and can be described and argued for to communicate to peers what choices have been made, why they have been made, and what their consequences are.

A database is a suitable technology to support data scopes, as it allows querying and storing (intermediate) results. We do not propose that a data scope should incorporate all data of a data cluster, as this would often require moving or duplicating entire, usually large data sets. This also does not mean that data scopes are limited to textual sources or metadata records, as the records in a database may contain links to any sort of data, including audiovisual materials, images and maps. For a data scope it is more important that it integrates data into a (research) view, by using data axes that should be queryable through an adequate interface. The interface itself might take many forms, depending on the requirements and the available resources. They range from a graphical user interface for querying and visualization, to API access and other forms of data import and export. In addition to querying, a data scope, has (access to) a chain of tools for elaborating data. The tools themselves are not essential for the data scope, and it is often a good idea to use external tools or a combination of tools for elaborating data, especially since specific tools tend to have a limited shelf life. The transformed data, however, should be part of the data axes and the resulting grid and be documented in the data scope. In this way, a data scope will store incremental data that are an expression of the sources and the research conducted upon them.

Data scopes take a technological form, as they operate within a data environment, but their objective is to mediate a research focus on a cluster of data using and documenting approaches of iterative modelling, linking data and classification.

To use a data scope, research questions should be translated to queries on a data cluster. Main research questions motivating the research are often not directly translatable to queries, but they can be broken down into separate, smaller questions that are more directly connected to specific queries and analyses.

Data scoping as a research method

Creating data scopes involves many decisions that are pertinent to the historical research question and affect interpretation and the creation of a historical narrative. Therefore, creating data scopes is part of the research process and methodology.

The activities of modeling, linking and classifying are connected to a researcher's assumptions and understanding of the data. The research process is inherently iterative in nature and there is no fixed order of steps. Each discovery of relevant narrative building blocks in the data cluster, or of relevant data sets to be added to the cluster, can trigger the researcher to reconsider choices in previous steps. It is easy to lose track of the many minute details and decisions along the way. To properly establish data scoping as a transparent research method, the tooling should support the researcher in keeping track of how a data scope is created. That is, it can automatically log which selection choices were made, which normalizations performed and which data sets were linked based on which criteria.

Each specific research question is thereby explicitly and transparently connected to a data scope on a data cluster. Each data scope should be accompanied by a description of explicit transformative steps that were taken to provide that research focus. A description of transformative steps that create a data scope can be published alongside research publications based on it. This allows more direct criticism on method and techniques in relation to data and research question. It opens up the "black" box to see what actually happened to the data, and how that affects interpretation. Others can peer through a researcher's data scope, which provides them with the same perspective. Reporting on the construction of a data scope should be able to address the questions: how does the data observed through a certain query on the data scope relate to the original source data? How can a scholar use knowledge of the construction of a data scope to make sense of the transformed data in the context of the research question and the source material? In communicating new insights, how should a scholar describe the data scope, so that others can perceive and understand these insights?

The development of large-scale humanities research infrastructures aims to provide scholars with tools for perform scholarly activities (Anderson et al. 2010), but these often attempt to offer user-friendly graphical interfaces over harmonized and linked data sets, based on the argument that humanities scholars lack the technical skills to do this themselves. But this

user friendliness often comes from hiding the many modelling choices, processes of selection, normalization and data cleaning, and parameter settings of search and visualization algorithms, which make the process opaque to the researchers and leaves them with less power to interpret and critique the sources, tools and data. Explicit descriptions of the provenance of a data scope could remove this impediment. Apart from making it easier to perform source and tool criticism, these descriptions also allow reuse, modification and extension to create new data scopes.

The critical analysis of data and the activities that transform it to make a data scope to answer research questions should all be seen as part of a process of digital source criticism. Source criticism has always been one of the key skills of the historian, that may be seen as a way to assess the research possibilities, perspectives and limitations of a given research corpus that forms the raw material from which a historical narrative is crafted. In the digital realm, traditional source criticism is still important, but it should be complemented by the critical examination of the data-specific issues that have become part of much historical research. Such an extension of historical source criticism is vital for doing history in the digital ag, because of the size of data, their multiple origins, the frequent involvement of other experts, and the many simplifications, abstractions and structuring that are necessary to use large-scale data to answer research questions. Data scopes are a methodological vehicle for digital source criticism.

Conclusion

Data Scopes are a concept of the interaction between researchers and their data in digital historical research. They go beyond the idea of a neutral and inert cluster of data that is the source for a researcher in which he or she can discern patterns and find information. A data scope uses an iterative process of selection of sources and data, modelling for research questions and defining data axes and subsequent normalization, linking and classifying data to make disparate data sources accessible for exploration, querying and comparison. Historians tend to view modeled data as a straightjacket for their research. Our view of data scopes is that they can be enabling under the provision that researchers invest effort to interact with their data and do not expect data clusters to yield ready-made answers. In interacting, they shape their own views of the data, that are rooted in and emanate from research questions. A data scope is designed to mediate the process of knowledge creation and representation as well as keep track of data elaborations and enhancements and in this way stays transparent. Data scopes are a concept and not a toolbox and may incorporate any number of data transformations researchers want to use. They should be flexible enough to adapt to the researchers' needs and open enough to incorporate new data sets.

In line with Edelstein (2017), we argue that data scopes do not revolutionize the field but offer a way to use what is already available in a more explicit and transparent way. It does not replace existing methodology but seeks to extend it with digital tools and incorporate heuristics into a digital environment that is too large to accommodate existing methodologies unaltered. We realize that many of the methods we propose and combine are already in wide use with researchers, data scientists and digital history groups. Still, they have never

been formulated as a consistent and coherent methodology or praxis. They are an extension of and complement to historical source criticism for the digital age. Tom Scheinfeldt (2008) points out the problematic but necessary change in values: "All of these things —collaborative encyclopedism, tool building, librarianship— fit uneasily into the standards of scholarship forged in the second half of the 20th century. Most committees for promotion and tenure, for example, must value single authorship and the big idea more highly than collaborative work and methodological or disciplinary contribution. Even historians find it hard to internalize the fact that their own norms and values have and will again change over time." Methodological contributions are little valued and it may be difficult to encourage scholars to invest more in understanding digital technologies as integral parts of research methodology as long as the default perception is that the “real research” happens after digital data has been cleaned, normalized and organized.There is still a separation between research and data handling. Our main point is that data interaction should be seen as an integral part of doing research. There should be no more room for the sentiment that after the ‘data stuff’ has been done, the researcher can start doing ‘real research’. The data stuff is real research.

Acknowledgements

We gratefully acknowledge that useful comments on an early draft by Victor de Boer, Marijke van Faassen and Max Kemman, and the valuable and constructive feedback from the reviewers.

References

Anderson, S., T. Blanke, and S. Dunn. 2010. Methodological commons: arts and humanities e-Science fundamentals. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 368(1925), 3779-3796.

Arguing with Digital History working group. 2017. Digital History and Argument, (white paper). Roy Rosenzweig Center for History and New Media (November 13, 2017): https://rrchnm.org/argument-white-paper/.

Ashkpour, A., A. Meroño-Peñuela, and K. Mandemakers. 2015. The Aggregate Dutch Historical Censuses. Historical Methods, 48:4, 230-245.

Besser, H. 2004. The Past, Present and Future of Digital Libraries. In A Companion to Digital Humanities (eds S. Schreibman, R. Siemens and J. Unsworth), Blackwell Publishing Ltd, Malden, MA, USA.

Blair, A.M. 2011. Too Much To Know: Managing Scholarly Information before the Modern Age. New Haven, Conn.
[available at http://raley.english.ucsb.edu/wp-content/uploads/Reading/Blair.pdf]

Blanke, T., and M. Hedges. 2013. Scholarly primitives: Building institutional infrastructure for humanities e-Science. Future Generation Computer Systems, 29:2, 654-661.

Boer, V. de, A. Meroño-Peñuela and N. Ockeloen. 2016. Linked Data for Digital History: Lessons Learned from Three Case Studies. In Historiografía digital: proyectos para almacenar y construir la Historia. Mirella Romero Recio and Mª Jesús Colmenero Ruiz (eds.) Anejos de la Revista de Historiografía. Universidad Carlos III de Madrid; (1 octubre 2016). ISBN 8416829012

Boer, V. de, M. van Rossum, J. Leinenga, R. Hoekstra. 2015. The Dutch Ships and Sailors Project. DHCommons Journal, 1.
http://dhcommons.org/journal/issue-1/dutch-ships-and-sailors-project

Brenninkmeijer, C., et al. 2012. Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision.
http://linkedscience.org/wp-content/uploads/2012/05/lisc2012_submission_8.pdf

Burke, T. 2011. How I Talk About Searching, Discovery and Research in Courses. May 9, 2011.
https://blogs.swarthmore.edu/burke/blog/2011/05/09/how-i-talk-about-searching-discovery-and-research-in-courses/

Edelstein et al., 2017 “Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. Historical Research in a Digital Age.” American Historical Review 122, 400-424

Erickson, A.T. 2013. Historical Research and the Problem of Categories: Reflections on 10,000 Digital Note Cards. In "Writing History in the Digital Age", Kristen Nawrotzki; Jack Dougherty, Digital Humanities Series, Published: Ann Arbor, MI: University of Michigan Press, pp. 133-145. DOI: http://dx.doi.org/10.3998/dh.12230987.0001.001

Fickers, F. 2012. Towards a New Digital Historicism? Doing History in the Age of Abundance. View journal, volume 1 (1). http://orbilu.uni.lu/bitstream/10993/7615/1/4-4-1-PB.pdf

Flanders, J., Jannidis, F. 2015 Data Modeling, in A New Companion to Digital Humanities (eds S. Schreibman, R. Siemens and J. Unsworth), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781118680605.ch16

Fry, B. 2007. Visualizing data: Exploring and explaining data with the processing environment. O'Reilly Media, Inc.. Chapter 1: The Seven Stages of Visualizing Data.
https://www.safaribooksonline.com/library/view/visualizing-data/9780596514556/ch01.html

Gibbs, F., and T. Owens. 2013. The Hermeneutics of Data and Historical Writing. In "Writing History in the Digital Age", Kristen Nawrotzki; Jack Dougherty, Digital Humanities Series,

Published: Ann Arbor, MI: University of Michigan Press, pp. 159-170. DOI: http://dx.doi.org/10.3998/dh.12230987.0001.001

van Faassen, M., R. Hoekstra, J. Ensor. 2015. Ruptured Life Courses : Institutional and Cultural Influences in Transnational Contexts. DH2015 Global Digital Humanities Conference Abstracts. 2015. http://dh2015.org/abstracts/xml/VAN_FAASSEN_Marijke_Ruptured_Life_Courses__Instit/VAN_FAASSEN_Marijke_Ruptured_Life_Courses__Institutiona.html

Gibbs, F., T. Owens. 2013. The Hermeneutics of Data and Historical Writing. In Kristen Nawrotzki; Jack Dougherty. Writing History in the Digital Age. University of Michigan Press, 2013. DOI: http://dx.doi.org/10.3998/dh.12230987.0001.001.

Glaser, B. G., and A.L. Strauss. 2009. The discovery of grounded theory: Strategies for qualitative research. Transaction publishers.

Graham, S., I. Milligan, and S. Weingart. 2015. Exploring Big Historical Data: The Historian's Macroscope. London, Imperial College Press, 2015, ISBN: 9781783266371.

Groth, P., Y. Gil, J. Cheney, and S. Miles. 2012. "Requirements for provenance on the web." International Journal of Digital Curation 7(1).

Guiliano, J. 2017. Toward a Praxis of Critical Digital Sport History. Journal of Sport History, Volume 44, Number 2, Summer 2017, pp. 146-159.

Guldi, J., and D. Armitage. 2014. The History Manifesto Cambridge: Cambridge University Press.

Hitchcock, T. 2013. Confronting the Digital - Or How Academic History Writing Lost the Plot. Cultural and Social History, Volume 10, Issue 1, pp. 9-23. https://doi.org/10.2752/147800413X13515292098070

Hoekstra, R. 2007. Correspondentie totaal. Patronen en trends in de briefwisseling van Willem van Oranje. In: Eef Dijkhof, Michel van Gent(eds), Uit diverse bronnen gelicht. Opstellen aangeboden aan Hans Smit ter gelegenheid van zijn vijfenzestigste verjaardag, The Hague: Instituut voor Nederlandse Geschiedenis, pp.117-131 https://www.researchgate.net/publication/283725888_Correspondentie_totaal_Patronen_en_trends_in_de_briefwisseling_van_Willem_van_Oranje

Huistra, H., and B. Melink. 2016. Phrasing history: Selecting sources in digital repositories. Historical Methods, 49:4, 220-229.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. and Gay, G., 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. ACM Transactions on Information Systems (TOIS), 25(2), p.7.

Klein, W., K. Zervanou, M. Koolen, P. van den Hooff, F. Wiering, W. Alink, and T. Pieters. 2017. Creating Time Capsules for Historical Research in the Early Modern Period: Reconstructing Trajectories of Plant Medicines. The 4th International Workshop on Computational History.

Lindblad, J.Th. 1995. Dutch entries in the pound-toll registers of Elbing, 1585-1700, The Hague: Instituut voor Nederlandse Geschiedenis available at http://resources.huygens.knaw.nl/retroboeken/elbing

McCarty, W., 2004. Modeling: a study in words and meanings. In A companion to digital humanities, (eds S. Schreibman, R. Siemens and J. Unsworth). Blackwell Publishing Ltd, Malden, MA, USA..

Ockeloen, N., A. Fokkens, S. ter Braake, P. Vossen, V. de Boer, G. Schreiber and S. Legêne. 2013. BiographyNet: Managing Provenance at multiple levels and from different perspectives. In: Proceedings of the Workshop on Linked Science (LiSC) at ISWC 2013, Sydney, Australia, October 2013.
http://linkedscience.org/wp-content/uploads/2013/04/paper7.pdf

Owens, T. 2014. Where to start? On Research Questions in The Digital Humanities.
http://www.trevorowens.org/2014/08/where-to-start-on-research-questions-in-the-digital-humanities/

Piersma, H., and K. Ribbens. 2011. Digital Historical Research: Context, Concepts and the Need for Reflection. BMGN - Low Countries Historical Review. 128(4), pp.78–102. DOI: http://doi.org/10.18352/bmgn-lchr.9352.
http://www.bmgn-lchr.nl/articles/abstract/10.18352/bmgn-lchr.9352/

Putnam L. 2016. The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast. American Historical Review, Volume 121, Number 2, pp. 377-402.

Scheinfeldt, T. 2008, Sunset for Ideology, Sunrise for Methodology? (Blogpost)
http://foundhistory.org/2008/03/sunset-for-ideology-sunrise-for-methodology/

Schmidt, B. 2016. Do Digital Humanists Need to Understand Algorithms? Debates in the Digital Humanities, 2016 Edition.

Schöch, C. (2013) Big? Smart? Clean? Messy? Data in the Humanities. Journal of Digital Humanities, 2013, 2 (3), pp.2-13

Sperberg-McQueen, M. 2004. Classification and its Structures. In A Companion to Digital Humanities, Blackwell Publishing Ltd., pp. 161-176.

Star, S. L. 1998. Grounded classification: Grounded theory and faceted classification. Library trends, 47(2), 218.

Traub, M.C., J. van Ossenbruggen. 2015. Workshop on Tool Criticism in the Digital Humanities. Workshop report. http://oai.cwi.nl/oai/asset/23500/23500D.pdf

Underwood, T. 2014. Theorizing Research Practices We Forgot to Theorize Twenty Years Ago. Representations, Vol. 127 No. 1, Summer 2014; (pp. 64-72) DOI: 10.1525/rep.2014.127.1.64.

de Valk, J.P. 1995. Ontsluiting van Egodocumenten, in: B.J.G.de Graaff en A.J. Veenendaal jr (red.), Bronontsluiting voor de negentiende en twintigste eeuw. Teksten van een symposium georganiseerd door het Instituut voor Nederlandse Geschiedenis, Den Haag, Instituut voor Nederlandse Geschiedenis, 1995, 37-41 http://resources.huygens.knaw.nl/pdf/MethodenEnTechnieken/Symposium.pdf

Zijdeman, R. 2016. Work In A Globalised World. Allocation Algorithm To Add Labour Relations To Digitised Census Data. ADHO Digital Humanities 2016. http://dh2016.adho.org/abstracts/306