From Serial Sources to Modeled Data
Changing Perspectives on Eighteenth-Century Court Records from French Pondicherry[1]

Anna Dönecke

A crucial point in historical research is the availability of sources, which is, especially for pre-modern times, not guaranteed. Hence, when I started my research on the French trading company in India during the eighteenth century, the issue of finding proper sources was a decisive point. However, as I wanted to focus on the jurisdiction in the French headquarter Pondicherry, it soon became clear that a dearth of sources would not pose a problem. The French administrators there produced an abundance of different documents, ranging from letters to the general directors in Paris and protocols of their meetings to a myriad of court records. Being enabled by the sources to delve deeply into the everyday proceedings of the jurisdictional field in Pondicherry soon proved to cut both ways. Although it may be presumptuous to speak of *big data* in comparison to the quantity of data natural scientists need to handle, the records provided a vast amount of historical data.[2] The task of processing all the information thus posed a major challenge, and at this point digital methods came into play in my humanistic research.

The best option to cope with the situation was to gather all available information in a relational database in a structured way which would also

1 I am very grateful to Antje Flüchter and Stephan Fasold for critical readings of the article.

2 For a discussion of *big data* and their handling in historical cultural studies cf. *Schmale, Wolfgang*, Big Data in den historischen Kulturwissenschaften, in: Wolfgang Schmale (ed.), Digital Humanities: Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität, Stuttgart: Franz Steiner, 2015, 137.

allow me to analyze them further. But to look at this in a merely output-oriented way and to conceive digital methods just as some kind of auxiliary tool for historians would be, as I want to show in this article, too short-sighted. Rather, their application itself or, in this case, the modeling of the database should be focused as an encounter of different ways of asking questions and handling source material which already can be productive. Like the digital humanist Willard McCarty puts it, the computational demand for complete explicitness and absolute consistency "effects a sea-change by forcing us to confront the radical difference between what we know and what we can specify computationally, leading to the epistemological question of *how we know what we know*."[3]

Taking these observations as a starting point, this article focuses rather on the *making* of the database for my project than the database itself. It takes a humanistic point of view and conceives the modeling of data as "a constructive and creative process"[4] starting with unwieldy information in a bulk of sources and ending with a formal model representing this information in a structured way. To do so, I focus on two questions while tracing the process of modeling the database for my project. On a general level, I ask what it means for a trained historian to apply digital methods in her research and point out some specific difficulties arising from historical sources. Furthermore, I examine the way in which the necessity of developing an abstract grid for the database which grasps all relevant information of the source retroactively affects how one conceives the sources.

I will begin with a brief depiction of my research project, followed by a description of the database, its specific requirements, and the reason for its choice. Then, I will track the modeling of my data in three sections: In the initial step, I reconsider the sources and derive a conceptual model from them, which is then transferred into a structure for the database, before I concern myself with the issue of importing data into the database.

3 *McCarty, Willard*, Humanities Computing, Basingstoke: Palgrave Macmillan, 2014, 25.

4 *Flanders, Julia/Jannidis, Fotis*, Data Modeling, in: John Unsworth/Raymond George Siemens/Susan Schreibman (eds.), A New Companion to Digital Humanities, Chichester, West Sussex, UK/Malden, MA, USA: Blackwell, 2016, 234.

Point of departure: intercultural jurisdiction in eighteenth-century Pondicherry

My research project[5] on the intercultural jurisdiction in Pondicherry draws on the observation that, upon its arrival at the Coromandel Coast, the *Compagnie des Indes Orientales* was by no means in the position to dictate its terms. Rather, it found itself in a multipolar constellation of power, even after acquiring Pondicherry in 1674 from Sher Khan Lodi, the region's ruler. Not only was it competing with other European powers present in South India but it was also dependent on the different regional rulers' benevolence and relying on the local society's cooperation to establish a flourishing trading post. In this situation, the jurisdiction in Pondicherry was a central field of interaction where members of the different groups encountered one another. Also, it was an arena in which conflicts between individuals were decided as well as power relations between the French and local groups negotiated.[6]

In my project, I examine this process of negotiation within and on the jurisdictional field with a particular focus on the agency of local groups in

5 I am conducting this research as PhD project within the framework of the Collaborative Research Center SFB 1288 "Practices of Comparing. Changing and Ordering the World", Bielefeld University, Germany, funded by the German Research Foundation (DFG), subproject B01 "Order in diversity: Practices of comparing in intercultural jurisdiction (17th–19th century)" lead by Antje Flüchter and Christina Brauner.

6 This perspective should by no means imply harmonic conditions or neglect asymmetric constellations in situations of cultural contact. As Christina Brauner and Antje Flüchter emphasize in their article on state-building in a transcultural context: "For rule and governance to be established in a functional manner in a contact zone, the different cultural (in this case mostly governmental) routines have to be made compatible. […] Negotiation (in a broad sense) […] can take place in the context of different power constellations, from more or less balanced relations to situations of striking power asymmetries." *Brauner, Christina/Flüchter, Antje*, Introduction: The Dimensions of Transcultural Statehood, in: Christina Brauner/Antje Flüchter (eds.), The Dimensions of Transcultural Statehood, Leipzig: Leipziger Universitätsverlag, 2015, 23. Cf. also *Flüchter, Antje*, Structures on the Move: Appropriating Technologies of Governance in a Transcultural Encounter, in: Antje Flüchter/Susan Richter (eds.), Structures on the Move: Technologies of Governance in Transcultural Encounter, Berlin/London: Springer, 2012, 1–27; *Parasher, Gauri*, Between Sari and Skirt: Legal Transculturality in Eighteenth-Century Pondicherry, in: Christina Brauner/Antje Flüchter (eds.), The Dimensions of Transcultural Statehood, Leipzig: Leipziger Universitätsverlag, 2015, 56–77.

the light of increasingly prominent attempts to expand French control in Pondicherry during the eighteenth century. Therefore, I call into question contemporary claims of the French to legal sovereignty in their Indian trading post and ask for historical shifts in this context during the century which is commonly identified as the beginning of European or British colonialism in India. I assume that historical actors relied on practices of comparing to maneuver themselves through the pluralistic situation and to handle its legal and cultural diversity, and that these practices thereby played an important role in establishing a legal order in Pondicherry.[7] In particular, I am interested in how the different legal authorities – French and 'indigenous' – interacted with each other. Moreover, I ask how the different actors – litigants and judges – handled the co-existence of local as well as French legal orders and institutions, and pay particular attention to their practices of comparing.[8]

To do so, I draw upon a wide range of sources. To a large proportion the material consists of different types of documents from the *Conseil Supérieur de Pondichéry*, the highest French court in India. Those are mainly court records, correspondence, and regulations regarding the jurisdiction in Pondicherry. To broaden the perspective, I also use ego documents, for example letters from the administrators, memoirs from *Compagnie* employees or travelogues. Furthermore, the records of the *Tribunal de la Chaudrie* are of special importance. The *Chaudrie* functioned as a court for all local groups in civil matters and became progressively controlled by the French authorities in the course of the eighteenth century.[9]

---

7 For more information on practices of comparing, cf. *Epple, Angelika/Erhart, Walter (eds.)*, Die Welt beobachten: Praktiken des Vergleichens, Frankfurt a. M.: Campus, 2015.

8 For a more elaborate account and preliminary results of my project, see my article *Dönecke, Anna*, 'Le chapeau ou la toque': Rechtliche Vielfalt und soziale Diversität in Pondichéry im 18. Jahrhundert, in: Christina Brauner/Antje Flüchter (eds.), Recht, Ordnung, Diversität, Bielefeld: transcript, forthcoming.

9 For an introduction to the legal landscape of Pondicherry in the eighteenth century see *Houllemare, Marie*, La justice française à Pondichéry au XVIIIe siècle, une justice en 'zone de contact', in: Éric Wenzel/Éric de Mari (eds.), Adapter le droit et rendre la justice aux colonies: Thémis outre-mer, XVIe-XIXe siècle, Dijon: Éditions Universitaires de Dijon, 2015, 147–157.

Choosing a digital method: modeling a relational database

Although it is clear by now that my engagement with Digital Humanities served a specific research purpose, it is nevertheless necessary to stress this point when talking about the selection of a proper method or tool. As Julia Flanders and Fotis Jannidis put it, researchers in comparison to archivists or librarians "typically concentrate on producing data that will be more specifically directed towards their own research needs"[10]. Consequently, the research need is the guiding principle for choosing a proper method and tool.

Since I am especially interested in cases involving locals as well as Frenchmen, the *Chaudrie* records take up a central role in my research and thus served as the starting point. A part of them was edited and published by Jean-Claude Bonnan in two volumes, roughly comprising 250 cases in total.[11] These records are available as PDF scans and could therefore be handled by means of digital methods. First of all, the scans were transformed into machine readable text via OCR.[12] The records follow the same structure; all provide certain information at a specific position in the text with standardized phrasing: A header with three lines was added in the course of the publication, indicating the date of the trial, the name of the litigants, and their social affiliation, respectively. Hence, it was possible to computationally extract specific information by using regular expressions and gather them in a comma-separated values (CSV) file. In the CSV file, each line represented a case and consisted of several comma-separated fields that contained different information on it. As a result, the data was organized in a structured but still linear way, following the order of its appearance in the sources, and the options to evaluate it were very limited. At this point, a database seemed the most suitable tool to work with the collected data and to enable me to even

10 *J. Flanders/F. Jannidis*, Data Modeling, 233.

11 *Bonnan, Jean-Claude* (ed.), Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816, 2 vols., Pondicherry: Institut français de Pondichéry, 1999. For more information on this, cf. *Menski, Werner*, Jean-Claude Bonnan, Jugements du tribunal de la chaudrie de Pondichéry 1766–1817, in: Indo-Iranian Journal 46 (2003), 369–371. Yet, Bonnan's edition only comprises a small part of the records. A lot more survived,partly due to copies that were made in the nineteenth century. They are nowadays hold in the *Archives nationales d'outre-mer* in Aix-en-Provence.

12 For this, the OCR pipeline of the INF project was used. For more information, cf. the contribution by Patrick Jentsch and Stephan Porada in this volume.

add further findings from readings of the *Chaudrie* or the *Conseil* records. Let me illustrate this in more detail by drawing a comparison.

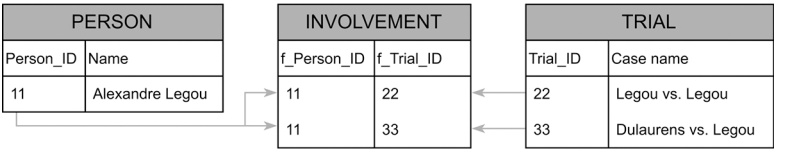
The comma-separated data can best be envisioned as a spreadsheet with each row containing information on a specific trial and the columns stating the date, the plaintiff's and the defendant's names, and potentially additional data like the judges' names, respectively. Obviously, this spreadsheet contained redundancies – judges were involved in more than one trial, of course, thus their names would appear multiple times on the spreadsheet. More importantly, there would only be a few options to manipulate the data. The lines could be arranged in a chronological order, sorted alphabetically by the names of the plaintiffs or by the number of occurrences of a specific name. But since the object of study in my research is foremost the *inter*action of different actors and authorities in Pondicherry's jurisdiction, more advanced options to record and interrogate the data were needed.

Quamen and Bath even choose this aspect as the defining point of databases: "A database, according to our definition, is a rigorously organized set of data whose informational patterns help us maximize the number of possible questions we can ask of it." They point out that "if you are embarking upon a project in which you will be actively engaging with your data, pushing its limits, and asking challenging questions of it – finding patterns, seeing changing dynamics over time, locating anomalies, looking for missing information – then you will need a database."[13] In order to stick to the notion of a spreadsheet, one can think of relational databases – the most common type – as multiple spreadsheets that are connected to one another. Apart from entity tables for, say, trials and persons, they also contain so-called junction tables that manage relationships between entities. In this example, they would match persons and trials together and, thus, capture information on the involvement of persons in trials. The mechanism by which this is done is identifying each set of data by a unique identifier (ID), called primary key, and using this ID, then called foreign key, to refer to it in other tables.

13 *Quamen, Harvey/Bath, John*, Databases, in: Constance Crompton/Richard J. Lane/Ray Siemens (eds.), Doing Digital Humanities: Practice, Training, Research, London/New York: Routledge, 2016, 181.

For example, as shown below, Alexandre Legou would be identified as 11 in the table 'person' while the inheritance dispute he was involved in would be put into the table 'trial' with the ID 22. Those sets of data would then be connected in a junction table called 'involvement' by matching 11 and 22 as foreign keys. If Alexandre Legou were involved in an additional dispute tagged with the key 33, one would create a new entry in the 'involvement' table using 11 again to refer to the very same data set in the 'person' table and connect it with 33. By following these primary-key-to-foreign-key links, a relational database manages relationships without creating redundancies.

Fig. 1: Example for tables connected via IDs



This distribution of data between multiple spreadsheets – the "atomization of data"[14] – and their interconnectedness increase the number of potential questions that can be posed to the data. Especially its ability to capture relationships between different entities made a relational database the most fitting tool for my research. It provided the possibility to organize my data in a way that allows me to pose questions that are directed at the entanglement of different actors and institutions on the legal field and to ask for diachronic changes. As a tool to build such a database, the database management system *FileMaker* was chosen because it not only provides the possibility to design a customized interface for possibly entering further data later on but also a tool to import CSV data.

From text to modeled data

As mentioned before, I want to focus on the modeling of data as a constructive process. That means neither the sources nor the data can be regarded as an adequate description of real historical entities but rather as an interpre-

14 *H. Quamen/J. Bath*, Databases, 184.

tation (of an interpretation) of them. The sources already represent a subjective perspective on historical events and by no means describe them as they 'really' took place. Moreover, the historian actively creates the set of sources he or she works with by choosing the most fitting for his research interest and leaving others out. This 'alienation' is even taken one step further when the scholar singles out specific aspects of the sources' abundance of possible information in the course of his or her research.

Taking this into account, modeling data can be understood as a "process of abstraction"[15] that starts with sources referencing historical entities and ends with a rather abstract or formal depiction that brings their information into line with computational requirements. In short, it is all about reducing and enclosing the 'unwieldiness' of data historians come across in their sources. During this process, it is crucial to keep in mind "that the formalized model determines which aspects of the subject will be computable and in what form".[16]

a. Reconsidering the sources: conceptual data modeling

Since the sources serve as the very basis not only for the automatic extraction of data but consequently also for the database, they are the starting point. So, the initial question to ask is: What information do they actually contain and which is relevant to my research purpose? The judicial records from Pondicherry show a wide range of different conflicts between individuals and contain a lot of detailed information about the trials. Yet, they describe a social situation with relatively fixed rules or roles. In each trial, there are a plaintiff and a defendant arguing about an issue that is finally decided upon by judges. In addition, court records even in the eighteenth century followed more or less specific rules or conventions. Those not only determined how they were written but also which information was included in the text. Hence, the records are structured similarly and contain the same categories of information. For example, very detailed information is given in most records about the written procedure, like the dates when requests were submitted by the litigants or the number with which the protocol was registered. Also, as different laws for different social groups existed, the social

15 *J. Flanders/F. Jannidis*, Data Modeling, 230.

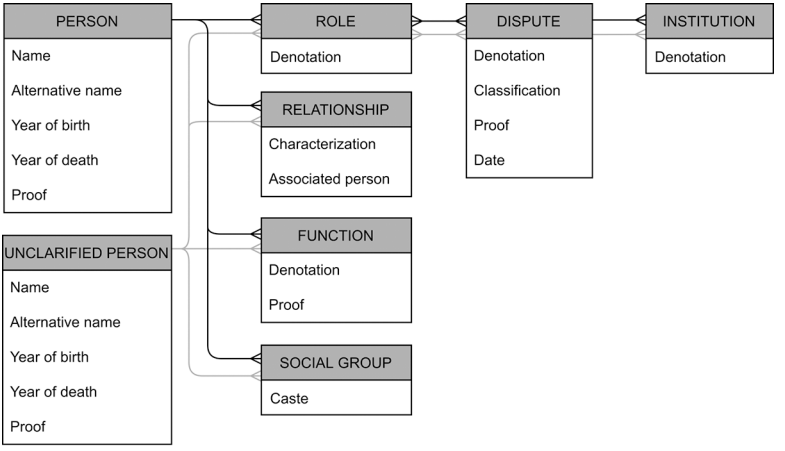
16 Ibid., 229.

affiliation or, to be precise, the ascribed social affiliation of the litigants was usually mentioned at some point in the record. As the disputes were often about family matters like inheritance issues, one often finds information about family relationships between the persons involved. Moreover, it was quite common for the *Chaudrie* to rely on local headmen, the so-called *chefs de c*aste, who functioned as arbitrators in smaller conflicts or as a court of appeal if the litigants were not satisfied with their arbitral decision. Consequently, its records often show traces of the interaction with them. Similarly, people could appeal to the *Conseil Supérieur* against the *Chaudrie*'s decisions. Finally, the *Chaudrie*'s records as well as those of the *Conseil* attained legal validity through the judges' signatures and thus, all end with the judges' names. However, not all of these details are immediately relevant to my research. As I am interested in Pondicherry's legal and social order and focus on the historical actors and their interaction, the following aspects could be singled out: Firstly, the personal information about the actors, i. e., their names, the denotation of their social affiliation, their role in the trial or their official function, their relationships to one another. Secondly, information about the trial itself is of interest, i. e., its date and the legal fora or institutions involved.

The next step is to convert these observations into a conceptual model. Here, a "purpose-oriented depiction is created that extracts a manageable amount of entities, attributes and relationships out of the plenitude of real world information".[17] The aim is to grasp the structure of the source's information and through this first abstraction prepare a basis for the actual database model. This is most commonly realized by means of an entity relationship diagram that can be understood as a visualized answer to the questions: What entities can be identified, what attributes do they have and how are they related to each other?

17 *Jannidis, Fotis*, Grundlagen der Datenmodellierung, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), Digital Humanities: Eine Einführung, Stuttgart: Metzler, 2017, 103. Original quote in German: "In der konzeptuellen Modellierung wird eine zweckgebundene Abbildung aus der Fülle realweltlicher Informationen auf eine überschaubare Menge von Entitäten, Attributen und Relationen erstellt."

Fig. 2: Conceptual model (together with Anna Maria Neubert from the INF team)



Each box in this diagram represents an entity with its attributes while the lines connecting the boxes are used to depict the types of relationships these entities share. The trident side of the line signifies 'many' while the straight end represents 'one'.

In the first place, there were seven entities to be distinguished in the court records: persons, the role they play in the trial, the dispute or trial itself, the institution or court, relationships the people had with one another, the office or function the persons held, and their social affiliation. The entity 'unclarified person' points towards specific aspects that one encounters when working with historical sources. Names were often spelled very differently, even if the same person was meant. Also, Indian converts were sometimes referred to by their Christian name, sometimes by their Indian name. Moreover, sons often bear the name of their fathers. As a result, it is sometimes – even if drawing on multiple additional sources – not possible to ultimately distinguish them for sure. For example, it is not possible to know whether Lazar and Tânappa Mudali[18] or Jacques Dulaurens who was

18 This example is taken from the description of the inheritance dispute that evolved after the death of Kanakarâya Mudali in 1747. See *Pillai, Ananda Ranga*, The Private Diary of Ananda Ranga Pillai: Dubash to Joseph François Dupleix […] A Record of Matters Political, Historical, Social, and Personal from 1736 to 1761, ed. and transl. by J. Frederick Price, Madras: Government Press, 1904, vol. 1, 310–375 and *Anonymus*, Arrêt du 20.03.1747, in:

appointed secretary at the *Conseil* in 1718 and Jacques Dulaurens appearing as a judge at the *Conseil* in a record from 1739[19] were two different persons or one and the same man.

Secondly, relevant attributes – again with regard to my research – were added to each entity. While attributes like 'name' or 'caste'[20] are derived more or less directly from the sources, other attributes were added due to the database's purpose. The attribute 'proof' is needed to meet the requirement of traceability. Aside from the necessity to keep track of the sources that have already been processed, one also needs to be able to give proof where the information was found. The attribute 'classification' is located at an analytic level, meaning that it is me as a scholar undertaking it. It already anticipates the possible need to sort out different types of disputes, for example inheritance or adoption issues, when analyzing the data in the end.

Thirdly, the diagram also represents the relationships the different entities share. A person usually has multiple relationships, being for example a father and a brother at the same time. Also, a person can have different functions: The judges of the *Chaudrie* were mostly members of the *Conseil Supérieur* at the same time. Although rather unusual, a person could be assigned to more than one social group. For instance, an Indian convert could be described as a Christian while still being seen as a member of the pariahs. A person could also be involved in more than one dispute and play different roles in each. For example, he or she could be the plaintiff in an inheritance dispute while being sued for not paying his or her debts in another trial. Lastly, there could be more than one institution involved in resolving a dispute. People could for example try to appeal to the *Conseil Supérieur* when

Gnagou Diagou (ed.), Arrêts du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 178–81.

19 *Anonymus*, Procès-verbal du 22.06.1718, in: Edmond Gaudart (ed.), Procès-verbaux des déliberations du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1913, vol. 1, 191–192 and *Anonymus*, Arrêt du 31.03.1739, in: Gnagou Diagou (ed.), Arrêts du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 58–59.

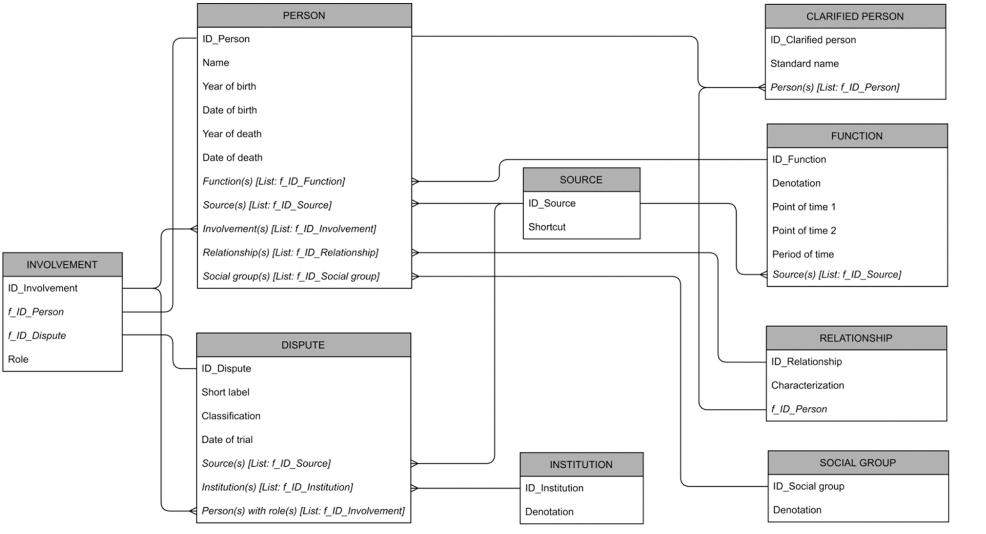
20 It should be noted here that the notion of a pan-Indian caste system as the defining feature of Indian society has been debated in the past decades. One widely shared result is that it was indeed implemented in the nineteenth century under the British Raj. Cf. for example *Dirks, Nicholas B.*, Castes of Mind: Colonialism and the Making of Modern India, Princeton, New Jersey: 2001. Nevertheless, the French frequently used the term caste in their records and classified the litigants by their (alleged) caste affiliation.

they were not satisfied with the *Chaudrie*'s decision or, more common, the *Chaudrie* referred a case to the caste chiefs.

b. Designing the database: logical data modeling

The next step on the path to a database is to transfer the entity relationship diagram, which represents the structure of the source's relevant data, into a final database structure: "While the *conceptual* model has its origins in structures of meaning, the emphasis of the *logical* model is on providing a structure for the data that allows the user to use a set of algorithms to answer questions of interest in relation to the data."[21] This, again, can be best visualized as a set of interconnected tables.

Fig. 3: Logical model (in cooperation with Silke Schwandt and Patrick Jentsch from the INF team)



At the center of this structure, there are the two master tables: 'personal data' and 'dispute'. These are the main entities identified before, all other entities are linked to either one or both of them in the entity relationship diagram. Persons and disputes are connected via IDs in a junction table called 'involve-

21 *J. Flanders/F. Jannidis*, Data Modeling, 231. Emphasis added.

ment' that captures which persons were involved in which disputes playing which role. Behind the two master tables reside several other tables containing additional information regarding 'personal data' and 'disputes'. They are connected to the two master tables through IDs.

As we have now reached the step which is about modeling the data in a way that finally meets computational requirements, specific challenges arise. Above all, a way to handle the uncertainty of data has to be found. The first problem already announced itself in the last step: the impossibility of always clearly identifying persons. In this case, it becomes necessary to allow potential redundancies in the 'personal data' table. Here, all persons appearing in the sources are gathered. For example, we would create a set of data for both, Lazar *and* Tânappa Mudali. If, by reading the source more carefully or consulting additional sources, I found out they were basically the same person, I would connect both sets of data to one another in the junction table 'clarified person'. Secondly, there is the possibility of incomplete or partial data. Sometimes, when looking for the date of birth or death of a person, all one can figure out is the year in which she or he was born or died. Finding the exact date is rather the exception. To cover both options, different fields were created in the 'personal data' table: year and date of birth or death. Thirdly, it is often unclear when exactly a person acquired or lost a function. In many cases, it is only possible to look for his (there were no female office holders) first and/or last appearance in the sources in a specific function. Also in this case, the fields were created accordingly: 'period of time', 'point of time 1', and 'point of time 2'.

Another challenge in this step was the removal of the 'logical' problems from the entity relationship diagram. As redundancies are unwanted in a database, the repetition of the 'proof' field throughout the entities or tables had to be avoided. Without a proper solution, the same source would have appeared numerous times, since a single record usually contains data on different entities and fields. By creating an extra table for sources and tagging each of them with an ID, it became possible to refer to sources in the 'proof' fields across the different tables by using a foreign key. Lastly, there was a many-to-many-relationship in the diagram between 'role' and 'dispute' because a person could be involved in different roles in more than one dispute. This had to be resolved into two one-to-many relationships stored in a junction table. Since a person could only obtain one role in a dispute, the former entity 'role' was transferred into an attribute and field in the junc-

tion table 'involvement', now denoting the role a person played in a particular trial.

The thus structured data allows for a series of challenging questions. One set is directed at proportions which can be found in the trials of the *Chaudrie*: What can be said about the distribution of litigants in terms of their social background? To what extent were intra-caste and inter-caste conflicts brought before the *Chaudrie*? What kinds of conflicts were mostly brought before the *Chaudrie*? In which amount of conflicts were the so-called caste chiefs involved? The second set of questions asks for diachronic developments and is designed to verify preconceived hypotheses. As mentioned before, one can observe from the sources that the French set out to expand and tighten their control over the jurisdiction in the second half of the eighteenth century. In this context, they also drew a sharper line between the *Conseil Supérieur*, which had been a court for the European population, and the *Chaudrie* by restricting access to the former for the local groups. The modeled data now allows reassessing the practical effects of this undertaking because it not only captures the interaction of different institutions in one and the same trial but also the dates of the trials. Hence, it is possible to ask if the frequency of appeals before the *Conseil Supérieur* against judgements from the *Chaudrie* changed over time. Furthermore, one is enabled to ask whether there was an increase in the involvement of lawyers in the disputes which could indicate a process of professionalization of legal counseling in the course of the century.

c. Question of practicability: data import

The decisive step after modeling the database is filling it with actual information from the sources. The idea was to use the automatically extracted CSV data and to import them into *FileMaker*. But before the actual import could start, a closer look at the CSV file was necessary. It soon becomes obvious that it contains a certain amount of errors. While the dates of the trials were identified correctly in all instances, some errors occurred regarding the litigants' names. They are normally separated by "c/" which stands for "contre" in the records. This made it possible to split them into parties: "participant_a" and "participant_b". As there are nonetheless a few deviations from the standardized phrasing that served as the basis for the regular expression, this could not always be handled correctly by the machine as one can see in the chart below.

Fig. 4: Screenshot of the CSV file showing exemplary data sets (opened in LibreOffice Calc)

	A	B	C	D	E	
1	**participant_a**	**participant_b**	**iso_date**	**date**	**participants**	**case_text**
2	Kader Sahib	Mirabaté	1766-11-11	Ma 11 novembre 1766	Kader Sahib c/ Mirabaté	Choulia. En droit islamiqu
3	François et	No correct split.	1766-12-09	Ma 9 décembre 1766	François et	Carachi Chrétiens, Topas.
4	Chinoudou	Dobascayen	1766-12-09	Ma 9 décembre 1766	Chinoudou c/ Dobascayen	Malabar païen et Malabar
5	Andipa (ou Andiapa)	Chavrimoutou	1766-12-19	Ve 19 décembre 1766	Andipa (ou Andiapa) c/ Chavrimoutou	Le droit de revendiquer ur
6	Tanapen	Levé	1767-01-23	Je 23 janvier 1767	Tanapen c/ Levé	Maure et Pally. Un père pe
7	Mahamadoussaib	Chinatamby	1768-04-11	Lu 11 avril 1768	Mahamadoussaib c/ Chinatamby	Musulman et Chetty. Une
8	Chinapen	Savamanden	1769-03-02	Je 2 mars 1769	Chinapen c/ Savamanden	Le tribunal peut contraind
9	Cavqué Sinnapen	Sarmanden	1769-03-04	Sa 4 mars 1769	Cavqué Sinnapen c/ Sarmanden	La vente sur saisie d un ir
10	Tailamé	Calati	1769-04-04	Ma 4 avril 1769	Tailamé c/ Calati	Les frais d entretien d un
11	Rangapa Soncoupam	Chavraia	1769-04-28	Ve 28 avril 1769	Rangapa Soncoupam c/ Chavraia	Le tribunal autorise une é
12	Ariatal vve Arlapen	No correct split.	1773-03-23	Ma 23 mars 1773	Ariatal vve Arlapen	Pally, chrétien. La conditic

For example, there are cases in which the litigants are a group of people, like the one represented in row 3 where the original text says "François and Jacques Tarabellion c/ la soeur de Marie Perera veuve Luc Carachi".[22] Here, another problem is already indicated: Female litigants are often not explicitly named but only referred to by means of their relation to male family members. Sometimes a short remark about their family status, e. g., their widowhood, is added, as can be seen in the case of "Ariatal v[eu]v[e] Arlapen" in the last row. This example reveals another particular problem: The lack of information on the second party of the dispute in the header or the absence of the "c/" which also leads to the output "No correct split". Apart from that, there are more random deviations like in the record represented in line 5 where there is expressed insecurity about the spelling of the litigant's name: "Andipa (ou Andiapa)". However, the overall error rate is rather low, and the correct information is readily accessible in the records' header. The problem of corrupt data could thus be solved without expending too much effort by manually post processing the file.

At least in theory and as initially planned, these data could now be easily imported into the *FileMaker* database. However, during this practice, one is confronted with the complicated issue of matching the CSV data to the database structure. It has been clear from the outset that some tables like 'clari-

22 *Anonymus*, Tarabellion c/veuve Luc Carachi, 09.12.1766, in: Jean-Claude Bonnan (ed.), Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816, 2 vols., Pondicherry: Institut français de Pondichéry, 1999, vol. 1, 3.

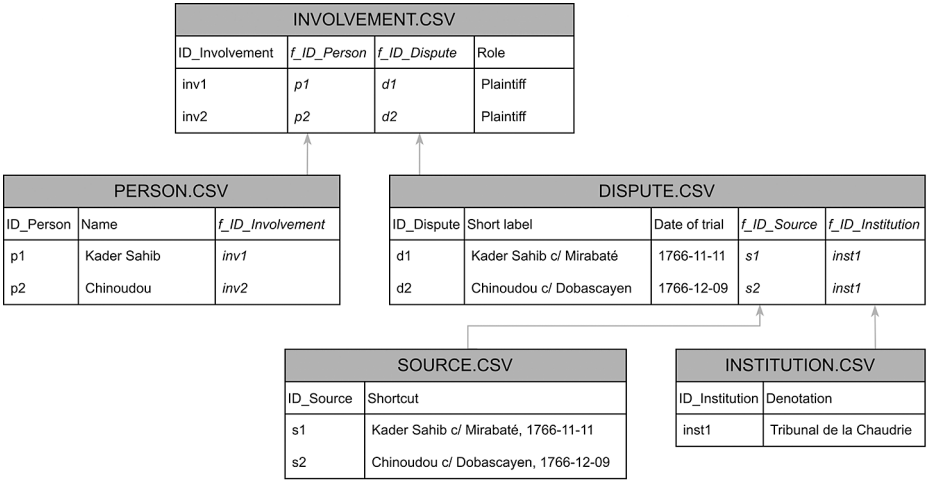
fied person' and 'relationship', and fields such as 'date/year of birth/death' in the database can only be filled manually while carefully reading the sources. This, unfortunately, also concerns the 'social group' of the litigants. Other than expected, the respective data could not be extracted from the records automatically because of their high degree of variation in this respect.[23] But most notably, there were several fields whose required information could be easily derived or generated from the CSV file but was not explicitly stored in it and thus could not be imported without further editing. Revisiting the distinction between information on entities and relationships, one can distinguish two different though closely related difficulties and approaches to solve them.

Firstly, entity data was missing in the CSV. With respect to the 'institution' table and its 'denotation' field, this could be solved easily because all evaluated records report on trials that took place before the *Tribunal de la Chaudrie*. Hence, a column entitled 'institution' was added to the CSV file, and its lines were all filled with "Tribunal de la Chaudrie". The values for the database's table 'source' could also be generated within the CSV file. The 'shortcut' field needed to distinctly reference the sources. As the records are sorted chronologically in the edition and there is not more than one trial with the same litigants and the same date, one could simply merge the values 'participants' and 'iso_date' in a new column. For example, the dispute between Kader Sahib and Mirabaté which took place on November 11th 1766 would be referenced as "Kader Sahib c/ Mirabaté, 1766-11-11". A bit more challenging but still manageable was the creation of values for the 'role' field in the junction table 'involvement'. This field refers to the particular role of persons involved in a specific dispute. As the plaintiff is always mentioned before the defendant, one can derive this information from the 'participants' as well as from the 'participant_a' and participant_b' values. Two new columns had to be added to represent these pieces of information in a form that is interpretable by *FileMaker*. In the first column entitled 'names', the values from 'participants_a' and 'participants_b' were compiled underneath each other. The second column named 'role' was accordingly filled with "Plaintiff" and "Defendant".

23 There is a not too small number of records where the social affiliation is not mentioned in the header. Also, the parties could both belong to the same social group or to different groups. Sometimes, even a single person can be ascribed to different groups as mentioned before.

Secondly, as described above, the defining advantage of a database is its ability to manage relationships between different entities, and the CSV file does not meet the corresponding requirements. To illustrate the problem in more detail, the notion of a database as a set of interconnected entity and junction tables can once again be useful. The CSV file, on the contrary, can be viewed as a mere entity table. While one knows that the values in one line all refer to the same trial, there is no explicit information at hand about this relationship. Consequently, when it comes to the *FileMaker* import, no values are provided for the junction tables which connect the different entities via IDs. Thus, respective IDs need to be added to the CSV file in order to represent the relationships between the different entities as defined in the logical model. Therefore, new CSV files corresponding to the tables defined in the logical model were created on the basis of the original CSV file. The table's field here became columns that were successively filled with the extracted (and corrected) data. Most importantly, an identifier for each set of data was generated (ID_Involvement, ID_Person, ID_Dispute, ID_Source, ID_Institution). These IDs were then copied to the other tables' corresponding columns and subsequently functioned as foreign IDs as illustrated below.

Fig. 5: Depiction of the different CSV files referencing one another by IDs



After this had been accomplished, there was explicit information on the different relationships between the entities in the form of IDs which could be interpreted by *FileMaker*. The CSV data finally matched the database structure which rendered it ready for its import.

Conclusion

The emphasis of this article was put on the way towards a historical database, seizing on a central observation from Willard McCarty that "models of whatever kind are far less important to the digital humanities than *modelling*. Modelling is crucial."[24] The main point of interest was the examination of the application of digital methods in historical research from a humanistic point of view.

It has become clear that through successively 'atomizing' the information from the court records into entities with attributes and relationships between one another, the sources' information could finally be transferred into a data structure that could be surveyed computationally. Also, some practical difficulties were laid bare in the process. Firstly, the uncertainty or partiality of data historians are used to work with emerged as a problem when facing computational demands. This could be solved by incorporating those factors into the data structure. Secondly, the approach to automatically import the data from the sources into the database turned out to bear unanticipated. Besides the need to manually correct corrupt data, the diverging structures of the CSV data and the database posed a challenge: The requirements for the data to be imported rise with the complexity of the evaluation potential of the database. On the one hand, the automatic extraction of data lightened the workload to a great extent; on the other hand, its results could only be used further if one was able and willing to invest further time in its post processing. At this point, one might see the limits of digitally assessing texts. But, revisiting the initial question of handling 'big data' posed in the introduction, what comes to the fore is, on the contrary, the potential of digital methods for handling a large corpus of serial sources.

Apart from this, the modeling of my data also forced me to actively engage with my source material in a specific way. Adapting my perception of

24 *McCarty, Willard*, What's going on?, in: Literary and Linguistic Computing 23 (2008), 254.

the sources to the requirements of data modeling meant to not look at them as individual disputes but rather to search for features they shared. Consequently, trials came to the fore as situations in which people and institutions interact with each other under relatively fixed rules. Bringing those configurations into focus also encompasses a particular attention for deviations one may not have seen otherwise. This tension brought to awareness the crucial question of how one conceptualizes the legal field and thereby became a reminder of one's own spatial and temporal embeddedness, one's 'Standortgebundenheit': As a twenty-first-century historian one tends to view the jurisdiction as a field thoroughly regulated by the state. But in eighteenth-century Pondicherry it was, in fact, a highly dynamic field where the trading company admittedly claimed jurisdictional power but rules for the everyday co-existence of the different groups were nonetheless subject to negotiations and were constantly on the move.

Moreover, this thinking in terms of abstract patterns and structures is a way to engage with the material that differs from the more usual perspective of a historian. As a hermeneutically working scholar one rather uses the material as a set of sources providing information about different disputes between historical actors which may give some indication of more general proceedings and, in a synopsis, of historical shifts. Appropriating the sources to a database's specifications, on the contrary, entails the necessity to develop *one* grid that can be applied to *all* sources. This means the historian from the outset has to radically detach her attention from the individual case and to redirect it to a level that as a rule encompasses more than one case and is oriented towards a formal description of all possible varieties.

In the end, it is exactly this change of perspective and the encounter of – from a historian's view – unaccustomed ways of looking at material that brings to awareness one's own premises and that renders the modeling of data enriching even if the actual data is not analyzed.

Bibliography

Anonymus, Arrêt du 31.03.1739, in: Gnagou Diagou (ed.), Arrêts du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 58–59.

Anonymus, Arrêt du 20.03.1747, in: Gnagou Diagou (ed.), Arrêts du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1935, vol. 1, 178–181.

Anonymus, Procès-verbal du 22.06.1718, in: Edmond Gaudart (ed.), Procès-verbaux des déliberations du Conseil Supérieur de Pondichéry, Pondicherry: Société de l'histoire de l'Inde française, 1913, vol. 1, 191–192.

Anonymus, Tarabellion c/ veuve Luc Carachi, 09.12.1766, in: Jean-Claude Bonnan (ed.), Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816, 2 vols., Pondicherry: Institut français de Pondichéry, 1999, vol. 1, 3.

Bonnan, Jean-Claude (ed.), Jugements du Tribunal de la Chaudrie de Pondichéry. 1766–1816, 2 vols., Pondicherry: Institut français de Pondichéry, 1999.

Brauner, Christina/Flüchter, Antje, Introduction: The Dimensions of Transcultural Statehood, in: Christina Brauner/Antje Flüchter (eds.), The Dimensions of Transcultural Statehood, Leipzig: Leipziger Universitätsverlag, 2015, 7–26.

Dirks, Nicholas B., Castes of Mind: Colonialism and the Making of Modern India, Princeton, New Jersey: 2001.

Dönecke, Anna, 'Le chapeau ou la toque': Rechtliche Vielfalt und soziale Diversität in Pondichéry im 18. Jahrhundert, in: Christina Brauner/Antje Flüchter (eds.), Recht, Ordnung, Diversität, Bielefeld: transcript, forthcoming.

Epple, Angelika/Erhart, Walter (eds.), Die Welt beobachten: Praktiken des Vergleichens, Frankfurt a. M.: Campus, 2015.

Flanders, Julia/Jannidis, Fotis, Data Modeling, in: John Unsworth/Raymond George Siemens/Susan Schreibman (eds.), A New Companion to Digital Humanities, Chichester, West Sussex, UK/Malden, MA, USA: Blackwell, 2016, 229–237.

Flüchter, Antje, Structures on the Move: Appropriating Technologies of Governance in a Transcultural Encounter, in: Antje Flüchter/Susan Richter (eds.), Structures on the Move: Technologies of Governance in Transcultural Encounter, Berlin/London: Springer, 2012, 1–27.

Houllemare, Marie, La justice française à Pondichéry au XVIIIe siècle, une justice en 'zone de contact', in: Éric Wenzel/Éric de Mari (eds.), Adapter le droit et rendre la justice aux colonies: Thémis outre-mer, XVIe-XIXe siècle, Dijon: Éditions Universitaires de Dijon, 2015, 147–157.

Jannidis, Fotis, Grundlagen der Datenmodellierung, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), Digital Humanities: Eine Einführung, Stuttgart: Metzler, 2017, 99–108.

McCarty, Willard, What's going on?, in: Literary and Linguistic Computing 23 (2008), 253–261.

McCarty, Willard, Humanities Computing, Basingstoke: Palgrave Macmillan, 2014.

Menski, Werner, Jean-Claude Bonnan, Jugements du tribunal de la chaudrie de Pondichéry 1766–1817, in: Indo-Iranian Journal 46 (2003), 369–371.

Parasher, Gauri, Between Sari and Skirt: Legal Transculturality in Eighteenth-Century Pondicherry, in: Christina Brauner/Antje Flüchter (eds.), The Dimensions of Transcultural Statehood, Leipzig: Leipziger Universitätsverlag, 2015, 56–77.

Pillai, Ananda Ranga, The Private Diary of Ananda Ranga Pillai: Dubash to Joseph François Dupleix […] A Record of Matters Political, Historical, Social, and Personal from 1736 to 1761, ed. and transl. by J. Frederick Price, Madras: Government Press, 1904.

Quamen, Harvey/Bath, John, Databases, in: Constance Crompton/Richard J. Lane/Ray Siemens (eds.), Doing Digital Humanities: Practice, Training, Research, London/New York: Routledge, 2016.

Schmale, Wolfgang, Big Data in den historischen Kulturwissenschaften, in: Wolfgang Schmale (ed.), Digital Humanities: Praktiken der Digitalisierung, der Dissemination und der Selbstreflexivität, Stuttgart: Franz Steiner, 2015.