The Digitisation of Handwritten Historical Documents: Database Solutions for Small Research Projects

Alexander Schunka
University of Stuttgart (Germany)

Introduction

This paper discusses the design and uses of database collection systems for personal and textual data in the historical sciences. The following considerations will apply to small- or medium-sized research projects, i.e., research enterprises of limited duration and resources, with a database used on one ordinary office PC for single-platform input. Based on case studies in early modern European history, carried out at the German Universities of Munich and Stuttgart, the following remarks demonstrate some possibilities of assembling, normalising and presenting data from heterogeneous, handwritten sources in relational databases.

The paper first gives a brief general outline of the state of digitisation in the German humanities, concerning the theoretical as well as practical aspects of the digital availability of historical information. Secondly, it focuses on methodological issues related to the digitisation of heterogeneous historical sources. Finally, it highlights the concrete possibilities and constraints in the digitisation of handwritten historical data and its presentation in databases as derived from two case studies. The first is a collection of biographical information on emigrants’ lives and careers in the German state of Saxony in the seventeenth and eighteenth centuries. The second case study is an edition of an early 18th-century German intellectual’s letters.

The paper, thus, discusses the advantages as well as the shortcomings of using conventional database systems in historical research. It is intended to give suggestions for similar small-sized database solutions in humanities research projects with limited personnel and technical resources.

Digital information in the historical sciences in Germany

In the German academic field, the digitisation of historical information, as well as the availability of digitised information on the Internet, has made immense progress in recent years. There are the big full-text databases such as ‘Projekt Gutenberg’, hosting some of the
most important texts from the dawn of the printing age onwards: Luther’s writings, for example, or Goethe’s works. There are online bibliographical databases deriving usually from printed bibliographies, such as VD17, the Jahresberichte für deutsche Geschichte, and many others. Most popular is the metasearch platform of regional library portals (KVK), enabling users to locate monographs in German and European libraries. Furthermore, the larger German state libraries have granted online access to their nineteenth- to twentieth-century book catalogues by scanning the mostly handwritten entries and making the digital images of their complete bibliographical data available on the net. Moreover, an initiative at the University of Erlangen-Nürnberg hosts retro-digitised bibliographical data of articles in German language historical journals – this task was fulfilled mainly by typing the bibliographical entries manually. There is not yet a German equivalent for the English-Language humanities database JSTOR, which contains the full texts of articles from some 600 different journals in the arts and humanities. The free-of-charge German historical newsletters, information portals, or mailing lists, with their steadily growing numbers of subscribers, draw their success largely from specific advantages of the internet such as hypertext linking, searchability, range, and speed.

The simple but crucial fact about the electronic dissemination of texts is, however, its digital availability, which makes the structuring or re-structuring of data necessary. The digitisation of larger and older historical volumes has usually been solved by either using the computer-edited texts and changing them into digital files, by presenting scanned image files of the book pages with limited searchability, or by typing the manuscripts into a computer whenever the script type is not yet automatically digitisable. Only recently has it been possible to digitise larger printed texts of older printing types, such as the nineteenth- and early twentieth-century German ‘Fraktur’ script, with the help of specially designed software.

However, the problem of assembling and displaying historical documents digitally has not been limited to printed materials. Even though the bigger archives, such as State archives, try to enlarge their web presentation, there are copyright concerns as well as technical and financial constraints which prevent them from presenting sources that are more primary online. In Germany, there are some well-documented attempts, such as the Web presentation of the Municipal archives of Duderstadt, the joint portal ‘Kalliope’ in Berlin, with its unpublished literary estates and autographs, and the abstracts of the Francke letters edition in Halle.
The goal of institutions like libraries and archives is usually not to do research of their own, but to host, preserve and make accessible historical documents for a wider, though mostly academic, public. This is the usual scope of library-based digitisation projects. The Bavarian State Library in Munich is particularly committed to digitisation projects funded by the German Research Foundation (DFG) and others. Their successes are impressive. However, academic research in Germany is usually not based in libraries but in universities. Here the limitations derive not only from technical or copyright questions. Apart from the big, long-term, library-based or academy-funded digitisation projects, research is being conducted at small-scale university institutes, sometimes reliant on the financial contributions of research foundations such as the Thyssen Foundation or the German Research Foundation, sometimes solely dependent on the ever-decreasing financial resources of the universities. The staffs of these enterprises often consist of a professor as the head of the project, one or two researchers and a few employed students. The time of funding is usually limited to three to six years, though sometimes, under very fortunate circumstances, it can go up to ten years. When it comes to considerations about how to present the results of research, plans for digitisation and digital publishing are likely to be rejected quickly, and the book will be preferred due to certain conservatism in the academic market. In many cases, further reasons are a lack of qualified personnel for digitisation, and inadequate technical resources. Mostly, researchers who are specialists in a certain field and know well how to deal with archival sources, lack an equal expertise in the technical implementation of digitisation. Moreover, the standing of digitisation work in the scientific community is sometimes still very low, depending on the aim of research, and despite ongoing efforts, mainly from younger scholars, to change this unsatisfactory situation.

**Methodological Considerations**

What is sometimes ignored in this context is the fact that digitisation of historical documents is not only a nice and practical way of preserving information, but enables the researcher to pose new questions to the material as well as to impose new methodological issues on the sources. Thus, digitising historical data can be part of the very natural process of scientific development and change in the humanities, concerning not only the form, but also the contents.

For a long time, assembling and analysing serial data has been a major issue in economic and social history. Innovative research with serial data has also been undertaken in cultural history, starting primarily with the methodology of the French ‘Annales’-school in the 1970s.
A particularly famous example is Michel Vovelle, who depicted the changing attitudes towards the Christian religion by analysing the language, *topoi* and stereotypes in ordinary people's wills.\(^{15}\) However, problems often arise from the non-statistical quality of the archival materials, which hardly seem to fit into any scheme of computerised data processing.\(^{16}\) This means that one of the most crucial issues is the choice and design of the means of data storage. It is important, first, to find a means of secure and long-term data storage that will ensure availability even some decades hence. Secondly, the presentation of the sources largely depends on the character of the material itself. For example, a database of biographical data requires a different design than the digitisation of letters, the analysis of grain prices, tax payments, etc. Finally, it is necessary to clarify the researcher’s interests, which will inevitably be mirrored in the design of the storage system. If, for instance, digitised texts of medieval or early modern chronicles are intended to serve as proofs of linguistic/semantic variation, other researchers may hardly use them to document the economic outline of medieval society. Thus, everybody dealing with digitisation needs to be aware that any kind of digitisation is already an interpretation from the digitiser’s point of view. This affects the historical interests, the methodological presuppositions, the choice of which kind of documents to digitise, how to digitise, and when to omit certain parts. Omissions or conjectures are sometimes necessary, due to source-related aspects such as lost, spoiled or unreadable material; or due to aspects of quantity of a *corpus* of sources, and in some cases even due to technical issues. In this respect, it must be stressed that even highlighting or italicising words in a printed edition is already part of the editor’s interpretation.\(^{17}\)

**Practical Issues: Two Case Studies**

In Germany, smaller research enterprises in the historical sciences often lack the institutionalised, technological, and human resources to deal with professional database solutions. This is equally true for the institutional background of the two projects mentioned. Considerations for both enterprises derive from the changing situation of the book and media market in Germany. Universities are currently facing serious financial pressure. Small university institutes are obliged to gain financial support for research from outside sources. At the same time, they need to show significant scientific results in a reasonable amount of time. Furthermore, the publishers’ policies lead to a massive reduction of efforts for printed scientific works, because these publications rarely sell out. This means that it can be financially difficult for universities to publish the results of their research, or even to buy their own researchers’ publications for the university libraries.\(^{18}\) It is, of course, one of the major goals of researchers to present their research to a wider audience. Nevertheless, the scientific
community in the humanities seems to be rather sceptical towards the concept of ‘open access’, although the situation might change within the next few years.19

The aforementioned considerations were essential for the outline of the two case studies in the digitisation of handwritten historical documents with which the author of this paper has been involved since the year 2000. The following remarks thus discuss some advantages, as well as disadvantages, of using conventional database systems such as Microsoft Access or Filemaker Pro together with the editing software TUSTEP.

The first digitisation project is a Filemaker-based database on the lives and itineraries of Bohemian immigrants in Saxony in the seventeenth century. It is jointly supervised by the University of Munich and the Saxonian State Archives of Dresden (Germany).20 It consists of biographical information on approximately 100,000 migrants, which is digitised in a database.21 This example highlights the practical issues and the problems of arranging a set of tables from sometimes inconsistent archival data. Furthermore, it provides insight into the possibilities of queries and the retrieval of information which was formerly unconnected and, therefore, almost inaccessible in the handwritten original copies. The project is due to be finished in 2007 and will be published either on CD-ROM or on the Internet.

The second case study is a recently started enterprise of editing the letters of an early eighteenth-century German intellectual, Daniel Ernst Jablonski, at the University of Stuttgart.22 Jablonski was part of a network of the most important European theological and political figures of his time.23 He wrote some 2,000 letters in at least six different languages (German, English, Polish, Czech, Hebrew and Latin). The focus will be on arranging a database with information derived from the contents of the letters, as well as from the biographies and bibliographies of his writing partners. Besides this, the database will be part of an edition of the letters themselves. Problems arise in the ability to define and search for keywords compatible with the different languages, and from the spelling inconsistencies and peculiarities of the handwritten sources.

Although the historical sources digitised in the two databases are different, the case studies share many similarities.

The first and most obvious common feature is that both projects deal with material from one of the core periods of early modern European History: the early seventeenth to the early eighteenth century, a period of increasing importance of state building and changing
opportunities of communication and mobility. Thus, both projects aspire to give clearer insight into the founding principles of early modern society, not directly from the point of view of rulers and states, but from below.

The second common feature is that the digitised material is of non-statistical quality. On the one hand, it consists of biographical information of ordinary Bohemian/Czech migrants who left their home country for religious, social and economic reasons, and who had to negotiate the opportunities for settling and integration in the bordering German states later on. On the other hand, it is the information in and around the letters of a multinational, multilingual intellectual based in Berlin. Thus, the original material is not comparable to the ‘classical’ early modern -- almost statistical -- sources like grain price lists, tax lists, ship-passenger surveys and other documents that were structured on a quantifiable basis from the start. The basis of the digitisation projects mentioned is comparable to the documents used by historians of an ‘histoire serielle des mentalités’.

The third similarity of the two enterprises is one of limited resources. Both the letter project and the migrants’ biography project are based at small German University institutes: the biography project in Munich and the letter project in Stuttgart, both at the Chairs of early modern European History. This means that, in both cases, the projects lack professional computer equipment and an extensive crew to deal with digitisation. In practice, work is taken out in both cases with small-range but widespread database software, and with one professional employee for the data input, and possibly a couple of students working part-time.

The fourth point of similarity between the biography project and the letter project is that both treat questions of transnational mobility, communication across borders, and last but not least, of emerging European identities. On a more abstract level, these topics seem to play an important role not just in the humanities in Europe, but also in today’s debates among historians of the African continent.

However, there are certain differences between the project on migrants’ biographies and the letter project. The first one relates to the differences in their sources. Therefore, different strategies of preparing and normalising the material need to be applied to make it suitable for any database structure. Secondly, in the case of the migrant biographies, we expect to publish the database itself, thus giving the sources’ information a somewhat artificial, semi-statistical quality. It is supposed to be published together with digital pictures of characteristic parts of the originals. In the case of the letter project, we eventually aspire to publish the full
transcripts of the letter texts. Here the database serves as a storing and connecting link between the information in and around the letters.\textsuperscript{26}

Both projects will continue for the next few years. The following remarks, therefore, serve only as an interim report. Still, it is worthwhile to mention certain considerations before setting off, and to look at the way things stand at the moment. For each project, it will be necessary to take a closer look at three aspects, namely, the quality and the peculiarities of the raw materials, the ways of normalising them in order to gain and retrieve digitisable data, and finally, questions of publication.\textsuperscript{27}

**The digitisation of the Bergmann immigrant collection**

The material basis of the migrants' database is a huge, handwritten collection. It includes information about approximately 100,000 people who left the Habsburg states, especially the Czech lands, and went to the bordering German areas in the seventeenth century.\textsuperscript{28} However, the present collection is itself already a selection of the original archival sources. The material was collected in its present form, transcribed from several archives and written down by the teacher Alwin Bergmann from Dresden during the first 40 years of the twentieth century. He managed to produce about 60 large, handwritten volumes. The present-day order of the so-called Bergmann Collection follows two main criteria, ‘names’ and ‘places’, and below that, refers to the migrants’ status (for example, nobility, clergy, craftsmen, agricultural workers, and others). Furthermore, it is alphabetised within the mentioned sections according to the migrants’ surnames. Each biographical entry, then, usually consists of some ‘hard’ data such as place and date of birth, death, places visited, the migrant’s itinerary, some quotations from archival sources or from printed literature, and citations of the material used. Still, the entries are very inconsistent because it was not always possible for the teacher Bergmann to trace back all the facts and dates. Owing to the state of preservation of early modern European historical sources, this would be impossible in some cases even for today’s professional historians, with more refined methods than Bergmann had in his time. Thus, the entries are different in length and content, although there are some regularities – which, admittedly, will look even more regular in a database structure. The definite value of the collection is its enormous size, but the information would be much more valuable if it were properly structured and connected. Before we started with the digitisation of the Bergmann migrants’ collection, it was hardly possible to raise questions on migrational peak periods, marriage patterns, baptisms, kinship, reverse migration, migrants’ networks, economic mobility and many other issues.
Despite some inconsistencies deriving either from the archival sources or from mistakes on Bergmann’s side, there are great opportunities to retrieve the socio-biographical information mentioned. Thus, it was decided to digitise the main parts of the collection with one employee and a small but functional database solution. The employee is not a professional historian. The input of historical information into a computer does not necessarily require a historian, although there are obstacles: Anyone who deals with the material needs to gain some understanding of the historical circumstances; he or she needs to be sensitive towards the raw material, and towards fitting it into a rather inflexible database structure; and he or she needs to be able to read the material and its ancient script in the first place. When it came to choosing the software, we decided to work with the database programme Filemaker Pro because of its stability, its opportunities for web presentation, and because it was fairly easy to use. The database system consists of 11 relational tables, structured primarily in accordance with the design of the written collection. The current amount of biographical data is approximately 30,000 records. There are about 2,000 records each for additional archival sources and for literature. The whole database is about 10 Megabytes in size, and thus easy to handle from one platform and not necessarily dependent on more equipment than a personal computer.

The digitisation project will be due in 2007. By then, we expect it to present the biographical information of about 100,000 people, based on biographical facts such as dates, places, kinship, and so on. Multiple spellings and different writings of names, for example in Czech and German, presented a problem. We decided to use alias entries for each different writing. In any search routine, these double records can easily be filtered out automatically. Besides, we plan to connect some characteristic entries with digital photographs of the original texts.

What is currently being discussed is the means of publication, together with our partner, the State archives of the German state of Saxony in Dresden, where the Bergmann migrants' collection and its digitisation are located. The original idea was to publish it on the Internet. However, the major archives in Germany do not yet favour publication of archival sources on the net. Thus, it looks like the most likely output will be a CD-ROM together with a tiny booklet, although that may require switching to a different software solution. At present, if we receive queries for biographical information on certain migrants from other researchers, either professional historians or semi-professional genealogists, we send them screen shots of the relevant biographical entries, quotable by primary key.
What was most striking, in the initial phase, was how carefully we, as researchers, had to choose which kind of data might be possible to digitise. A full-text digitisation would have been possible perhaps only as scanned image files, which would not have enabled the user to connect the information of different entries. However, the aim of the researchers was exactly that: the connection of hitherto unconnected bits of information. This is why it was necessary to determine which digitisable information fitted into a database. The original sources will therefore not be obsolete, but the results of the database queries will in certain respects go far beyond them. The biggest advantage is the connection of different, multifaceted migrant biographies either in personal networks or in statistical rows. In any case, there are deficiencies in the original sources, which cannot be evened out by any attempt at digitisation: They are incompletely preserved, scattered among many German archives and brought together by the teacher Bergmann over 40 years of his life. Though the material may sometimes be inaccurate or incomplete, it will hopefully still be possible to get closer to early modern social realities of mobility and communication.

We always needed to be aware that any kind of choice, amendment or omission in a text is already an interpretation. Still, we hope users will be able to articulate searches and formulate questions according to their own needs, so they can use the database for manipulation and analysis rather than just as a storage media. Thus, the migrants’ database is a small enterprise, but it will be valuable beyond its regional context.

The digitisation of Daniel Ernst Jablonski’s letters

The second case study has only recently left the planning stage. It is the database of information concerning the letters of a seventeenth- and eighteenth-century intellectual named Daniel Ernst Jablonski.

Jablonski was a contemporary of Gottfried Wilhelm Leibniz, the famous German philosopher, and also his friend and writing partner. The two thinkers were founders of a Prussian-German national academy of arts and sciences. They promoted science and humanities in their time, and Jablonski, in particular, was engaged in foreign politics, church politics and religious matters in Germany and Europe. Born in Poland, he had studied in England before going to the Brandenburg-Prussian king’s court, where he worked as the country’s court preacher, equivalent to a bishop. Though less well-known than Leibniz, he was a passionate letter-writer, with contacts reaching from America, England, and via the German states, as far as
Transylvania. Jablonski, being a multilingual cosmopolitan, wrote mainly in Latin, but also in German, Czech, Polish, English and Hebrew.

Jablonski’s letters are scattered all over Europe, as no consistent estate exists. Maybe 1,500 letters are already known; many more are not. From other sources such as inventories, or the estates of his contemporaries, we hope to be able to find altogether around 1,500 letters more, from and to Jablonski. The goal of the enterprise is therefore threefold: The letters must first be found and collected, which can be quite difficult, especially in East European archives with sometimes insufficient documentation; secondly, it is necessary to retrieve as much information as possible about each letter, such as contents, biographical information of the sender/recipient other than Jablonski, historical context, archival sources and literature relating to the letter; third, we hope to publish a (preferably) complete edition of the letters to and from Jablonski.

The whole project is scheduled to take ten to twenty years, if the financing is secured. The procedure is much smaller in scale than the enterprise of publishing the works and letters of Gottfried Wilhelm Leibniz at the Berlin Brandenburg Academy of Sciences in Potsdam. Still, the Leibniz Letters project serves as a model for the Jablonski edition, where similar sources and problems will be dealt with, and similar software solutions will be preferred. Unlike the case of the migrants’ database, the goal of the Jablonski project is the publication, not of a database, but of printed volumes, maybe together with a CD. However, the core of the project will be a database on the letter information.

Compared to the migrants’ database, the Jablonski database's design can be freer, in a way. There are no such seemingly clear-cut structures in the raw material as developed by Bergmann, but there are the letters and the information surrounding them. The work will be done with Microsoft Access. The technical details, such as tables and relationships, are more or less similar to those in the Filemaker system. The main tables will consist of ‘letters’ with abstracts of the letter contents, ‘persons’, ‘places’, and also of secondary information such as ‘archives’, ‘literature’ and so on. This database will be connected to the editing software TUSTEP, which has been designed especially for editing and publishing editions of primary sources. It is probably the most widely used software for this purpose in the field of the humanities in Germany. TUSTEP, which was developed at the University of Tübingen, meets all the necessary requirements of a scientifically correct edition, from questions of layout to a lavish, multi-layered footnote system and the display of different languages and fonts. Furthermore, it makes full-text indexing fairly easy. One drawback is that during data
input it is not possible to see the layout on the screen, which requires some adjustment from the user accustomed to common word processing software. The Leibniz-letter-project has been working with TUSTEP for a long time.\textsuperscript{36}

With TUSTEP, it is possible to produce either postscript or PDF-files. We, therefore, intend to make some of the edited letters of Jablonski accessible to the scientific community via PDF – even before the full volumes are printed.

The information database will serve as the background for each of Jablonski’s letters, which is necessary and helpful for the editors as well as for users later on. The database contents, such as biographical data, letter abstracts, etc., might be published in the future in one way or another, together with the letters, although the database itself as a whole will probably not be published.

Finally, the display of the letters in the edited volumes will probably follow a chronological order, as any attempt towards a systematic categorisation seems much more questionable to the editors. However, this bears the risk that later findings of hitherto unknown letters might have to be published in a supplement volume. Therefore, there is also the possibility of publishing the letters on the net in the future, if there are no copyright problems, or on CD.

For reasons of data permanence and more flexibility towards new developments in computing and digital publication, it has now become widely accepted in editing enterprises to use software standards compatible with or convertible to the Extensible Markup Language, or XML. TUSTEP is XML-compliant software, and thus it would be technically possible to publish and store the Jablonski letters project on CD-ROM.

It is important in this context to mention that over the last few years, there have been several efforts to define standards for editing, publication and conservation of primary sources in the humanities. The German Research Foundation, which is the most important organisation to launch and subsidise editing projects in Germany, probably started the most forceful initiative to set thorough standards for digital publication. The standards are based on the so-called Dublin Core Metadata Initiative, whose derivations have become more and more refined due to their practical uses in recent years.\textsuperscript{37}

However, the final decision on how to publish the Jablonski letters, i.e., whether or not to publish them on CD, has not yet been made. This is because of the ambivalence of the
German book market towards so-called ‘hybrid’ publications such as books together with CDs. Publishing it just on CD might not be a good idea, since the German market for historical publications tends to be rather conservative. Despite the fears in the 1990s that the book will be fully replaced by the CD, both kinds of publication still exist more or less happily together, though sometimes bearing different contents and catering to different needs.

**Conclusion**

For the two digitising projects mentioned, the migrants' database and the Jablonski letters, it was essential to consider five main points which are likely to be of equal importance in one way or another for other digitisation enterprises of similar outline. First, we wanted to make use of modern technologies in order to widen our knowledge on questions highly topical in our field of research. In both cases, early modern mobility and communication were the issues and they constituted key phenomena of early modern European society. Second, we had to be aware of limited resources, whether in technical equipment or personnel. Third, it was hardly possible to make the original, archival sources fit into the technical frame. The sources made it necessary to proceed the other way around, adjusting the technical preconditions to the materials dealt with. Certain compromises were inevitable, based on the constant awareness that any edition is already a selection and an interpretation. Fourth, we try to keep the scope as wide as possible in order to give future researchers the opportunity to express their own interests towards the material presented in the databases. The digitised material should, therefore, be not just a collection of texts, but a tool for research. Finally, it is essential that the people dealing with the design of the databases have a certain amount of knowledge of the historical sources and facts, and, even more importantly, that the historians preparing the editions gain some understanding of the technical possibilities for digitisation. Academics from both fields must learn from each other in order to contribute to preserving the high importance, not just of the historical sciences, but also of the humanities in general in the age of globalisation and global communication.

**Endnotes**

1 http://gutenberg.spiegel.de/info/info.htm [12.11.2004].


3 http://www.ubka.uni-karlsruhe.de/kvk.html [12.11.2004]

4 See, for instance, the Sächsische Landes- und Universitätsbibliothek Dresden (SLUB), http://image.slub-dresden.de/de/index.htm [12.11.2004]; and the Württembergische
See, for instance, the German branch of the US-based H-Net family:
http://hsozkult.geschichte.hu-berlin.de/ [12.11.2004]; the newsletter of the ‘Virtual Library Geschichte’, http://www.vl-museen.de/info/ [12.11.2004]; the newsletter on newspaper articles concerning the German historical sciences,
http://www.nfhdata.de/premium/index.shtml [12.11.2004], or portals like ‘historicum.net’,


For instance, the most important historical German encyclopedia, the ‘Zedler’:


http://www.archive.geschichte.mpg.de/duderstadt/dud.htm [12.11.2004];

http://www.bsb-muenchen.de/digital.htm [12.11.2004].


See W. Bader, ‘Was ist die Text Encoding Initiative (TEI)?’, Kamzelak, R.,ed.,


On Jablonski, see Dalton, H., 1903, Daniel Ernst Jablonski. Eine preußische Hofpredigergestalt in Berlin vor zweihundert Jahren, Berlin: Martin Warneck. (The author of this paper is currently preparing a monograph on Jablonski’s intellectual relations between England and Prussia.)


A screenshot of the main form is available at http://www.uni-stuttgart.de/hi/fnz/jablonskidb.html [12.11.2004].


The figures were accurate for August 2004.


http://dublincore.org/ [12.11.2004]