

Digitizing historical information from hand-written sources in databases:

Possible solutions for small-size research projects

Paper on the CODESRIA conference on electronic publishing and dissemination, 1 - 2 September 2004, Dakar, Senegal

Alexander Schunka, University of Stuttgart (Germany)

First of all let me thank you very much for your kind invitation to this conference. I am very glad to be here – being a non-social scientist, a non computer-specialist or librarian, but an historian dealing with Central Europe. However, I do believe that the goals and interests of CODESRIA and of this conference in particular share a lot of similarities with the questions of electronic publishing now being raised at University departments in Central Europe.

In my paper I want to treat some advantages and shortcomings of assembling and publishing historical information in databases. My paper is based on experiences with two research projects. One is a biographical database of 17th century czech emigrants into Germany, the other is a collection of private letters of an early 18th century intellectual named Daniel Ernst Jablonski.

Considerations for both case studies derive from the changing situation of the book and media market in Germany. Recently, universities are facing serious financial pressure. Small university institutes are obliged to gain financial support for research from outside. 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 due to the lack of money, it can be financially tough for universities to publish their results of research or even to buy their own researchers' publications for the university libraries. Still, it is of course one of the major goals for the researchers to present their results to a wider audience. Nevertheless, the scientific community in the humanities seems to be rather sceptical towards concepts of „open access“, although the situation might change within the next few years. This is why it is at the moment not definitely decided how the two database projects are going to be published. I will get back to this point later on.

Although the historical sources we are treating seem to be quite different, the two case studies have a great deal in common:

The first and most obvious common feature is of course that both digitisation projects deal with material from one of the core periods of Early Modern European History, namely from the early 17th to the early 18th century, which is believed to be a period of increased state building problems and changing opportunities of communication and mobility. Thus, both projects aspire to give clearer insight on the functioning principles of Early Modern society, not directly from the point of view of the rulers and states, but from below. The second common feature is that the material being digitized is primarily of non-statistical quality. On the one hand, it consists of biographical information of ordinary Bohemian or czech migrants who left their home country mainly for religious, social and economic reasons, and who had to negotiate the opportunities of 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. He wrote in at least six different languages (German, English, Polish, Czech, Hebrew and Latin). So the original material is not comparable to the classical almost-statistical sources of that time - like grain prices, tax lists, ship-passenger surveys and other documents which have been structured on a quantifiable basis even for their contemporaries. It is rather comparable to the documents used by historians of an „histoire serielle des mentalités“ such as Michel Vovelle who, in his famous work, showed the

changing attitudes towards christian religion by analyzing the language, the topoi and stereotypes in ordinary peoples' wills. The third similarity of the two enterprises is one of lacking 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 digitizing crew. In practice, we are trying to implement both with small-range but wide-spread database software such as Microsoft Access and Filemaker Pro, and each with one professional employee for the data input, and possibly a couple of part-time working students. The fourth point of similarity between the biography project and the letter project seems to be that both treat questions of transnational mobility, communication across borders, and last but not least, of emerging European identities. If I see it correctly, and according to the latest issue of the CODESRIA-Bulletin, these topics apparently play an important role in today's debates among historians of the African continent, and I hope this will in part justify my invitation and the following remarks.

However, there are certain differences between the project on migrant's biographies and the letter-project. The first and foremost is that the material is simply different. This forces us to apply different strategies of preparing and normalizing the raw material in order to get it databased. The second point is partly owed to the different material as well. In the case of the migrant biographies, we hope to publish the database itself, thus reducing the sources' information to some semi-statistical qualities, preferably together with some digital pictures of characteristic parts of the originals. In the case of the letter project, we want to publish the full transcripts of the letter text, where the database serves as a storing and connecting link to the information in and about the letters.

Both of the projects will be still under-way for the next couple of years, but I would like to give you some impression about our considerations before setting off with them, and about the way things stand at the moment. Let me start with the migrant biographies, and then continue with the letter project. For each project I want to take a closer look at three aspects, namely the quality and the peculiarities of the raw materials, the ways of normalizing them in order to gain and retrieve digitizable data, and finally questions of publication.

The material basis of the migrant's database is a huge, hand-written collection. It includes information about approximately 100.000 people who left the Habsburg states, especially the Czech lands, and who went to the bordering German areas in the 17th century. However, the present collection is in itself already a selection of the original archival sources. The material was collected, transcribed from several archives and written down by a teacher from Dresden during the first 40 years of the 20th century. He managed to produce about 60 large, hand-written volumes. The present day order of the so-called Bergmann-Collection follows two main criteria, „Names“ and „places“, and below that, recurs on the migrants' status (for example nobility, clergy, craftsmen, agricultural workers and others). Furthermore, it is alphabetized within the mentioned sections according to the migrants' surnames. Each biographical entry, then, usually consists of some „hard“ data such as place, date of birth, death, places visited, the migrant's itinerary, some quotations from the sources or from printed literature, and citations of the literature 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. Due to the state of preservation of Early modern European historical sources, this would in some cases be impossible 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 look even more regular in a database structure. The value of the collection is its enormous size. The disadvantage is that the information would be much more valuable if it were properly structured and connected. Before we started with the digitization of the Bergmann migrant's collection, we were hardly able to raise

questions on migrational peak periods, marriage patterns, baptisms, kinship, reverse migration, migrants' networks, economic mobility and many other things.

Despite some inconsistencies which derive either from the archival sources or from mistakes on Bergmann's side, there are great opportunities to retrieve information of almost any kind. Thus, we decided to digitize the main parts of the collection with one employee and a small, but functional database solution. The employee is not a professional historian. I believe this is not necessary for the input of information into a computer, but there are some 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 on top of this, he needs to be able to read the material in the first place. We decided to work with the software Filemaker pro because of its stability, because of its opportunities for web presentation, and because it is easy to use. The database system consists of 11 relational tables, structured primarily in accordance to the design of the written collection. The current amount of the biographical data is at about 30.000 records, we have about 2000 records each for archival sources and for literature, and the whole database is about 10 Megabytes in size, thus easy to work at from one platform and not necessarily dependent on the latest equipment.

The digitization project began about 3 years ago and will probably take another three years. By then, we expect to present the biographical information of about 100.000 people, although based on the so-to-speak „harder“ facts such as dates, places, kinship and so on. Problems to be solved were multiple spellings and different writings of names, for example in Czech and German. We decided to use aliases for each different writing. In any search routine, these double records can easily be filtered out automatically. Besides, we are planning to connect some characteristic entries with digital photographs of the original texts.

We are currently discussing the means of publication, together with our partner, the State archives of the German state of Saxony in Dresden, where the Bergmann migrant's 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 a publication of archival sources on the net, due to questions of copyright misuse. Only some small German archives like the municipal archive in the city of Duderstadt made their documents accessible through the internet. Probably, things will change soon, even for the bigger archives. Still, it looks like the most likely way of publication will be a CD-ROM together with a tiny booklet, although we are not sure whether to stick to the Filemaker database then.

What struck us most in the starting period, was the fact that we needed to choose carefully which kind of data might be worth digitising. We needed to be aware that any kind of choice, amendment or omission in a text is already an interpretation. Still we hope that the users will be able to articulate searches and formulate questions according to their own needs. If we currently receive queries for biographical information on certain migrants from other researchers, either professional historians or semi-professional genealogists, we are sending them screen shots of the relevant biographical entries, quotable by primary key. But the biggest advantage to me seems to be the connections of different, multifaceted migrant biographies either in personal networks or in statistical rows. There are enough deficiencies in the original sources, scattered among many German archives and brought together by the teacher Bergmann in 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. Thus, the migrants' database is a small enterprise, but it might be valuable beyond the regional context.

Let me now turn to the second case study, which has only recently left the planning stage. It is the database on information concerning the letters of a 17th and 18th 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 sciences 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 he went to the Brandenburg-Prussian king's court where he worked as the country's court preacher, equivalent to a bishop. Still less well-known than Leibniz, he was a passionate letter writer, with contacts reaching from England via the German states as far as Transylvania. He wrote mainly in Latin, but also in German, Czech, Polish, English and Hebrew, being a multilingual cosmopolitan of his time.

Jablonski's letters are scattered all over Europe. Maybe 1.500 letters are already known; many more are not. From other sources such as inventories, or the estates of his contemporaries we guess that we will be able to find altogether around 2.000 letters more, mainly from, but also to Jablonski. The goal of the enterprise is therefore threefold: First, we need to find the letters which can be quite difficult, especially in East European archives without sufficient documentation; secondly we will try to retrieve as much information about each letter, such as the contents, biographical information of the recipient, historical context, archival sources and literature. Third, we want to publish preferably a complete edition of the letters to and from Jablonski.

The whole project will take 10 to 20 years, if the financing is secured. The procedure is similar but much smaller in scale than the research enterprise of publishing the works and letters of Gottfried Wilhelm Leibniz. From a technical point of view, the Jablonski-project aims to be its „little brother“, as we use approximately the same software but lack the personnel and the institutional resources. The goal is not the publication of a database but the publication of printed volumes, maybe together with a CD. Still the core of the project will be a database on the letter information.

Compared to the migrants database, we are in a way much freer with the database design. We do not have seemingly clear-cut structures in the raw material as developed by the teacher Bergmann, but we have letters – and the surrounding information. Here we are working with Microsoft Access, but the technical details such as relationships of the tables are of course 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 is 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, like questions of layout, and the display of different languages and fonts. Furthermore it makes the full text indexing fairly easy. One drawback is that you do not always see what you get on the screen, as in the Microsoft world. The Leibniz-letter-project has been working with TUSTEP for a long time.

With TUSTEP, we can produce either postscript or pdf-files, so we will be able 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 thus serve as the background for each of Jablonski's letters, which is necessary and helpful for the editors as well as for the users later on. The database contents might be published in the future in one way or another, together with the letters; the database itself might not be published as a whole though.

In the final edition, the letters will probably be ordered chronologically, as any attempt at a systematic order seems much more questionable. However, this bears the risk that later findings of hitherto unknown letters might have to be published in a supplement volume.

Therefore, we will also consider publishing them on the net (if there are no copyright problems) or on CD.

During the last few years, there have been several efforts to define certain standards for editing, publication and conservation of primary sources in the German humanities. The German Research Foundation, which is the most important organisation to launch and subsidize editing projects, probably started the most forceful initiative to define standards for digital publication. For reasons of data permanence and of more flexibility towards new developments in computing and digital publication, it has now become widely accepted to use software standards compatible or convertible to the Extensible Markup Language XML. TUSTEP is a software compliant to the XML standards, and thus it would be possible to publish and store the Jablonski letters project on CD-ROM.

However, the final decision on how to publish the Jablonski letters, or of publishing them straight to CD has not yet been made. This is because of the reluctance of the German book market towards so-called „hybrid“ publications such as Book together with CD. Publishing it just on CD might not be a good idea either, since the German market for historical publications tends to be rather conservative, if I see it correctly. Despite the fears in the 1990es, that the book will be fully replaced by the CD, both kinds of publication still exist more or less happily together, although sometimes containing different contents and catering for different needs.

To conclude: For the two digitizing projects mentioned, the migrant's database and the Jablonski Letters, we had to consider five main points. First, we wanted to make use of modern technologies in order to widen our knowledge on Early modern mobility and communication, and thus, on key phenomena of Early modern European society. Second, we had to be aware of limited resources, be it the technical equipment or the people who deal with it. Third, we could hardly make the original, archival sources fit into the technical frame but rather the other way around - although there sometimes needed to be certain compromises, based on the constant awareness of the fact that any edition is already a selection and an interpretation. Fourth then, we were trying to keep the scope as wide as possible in order to give future researchers the opportunity to express their own questions towards the material presented in the databases. Fifth and finally, it seems to be 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 important, that the historians preparing the edition, gain some understanding of the technical possibilities for digitisation. Since I come from the side of the historical sources and not from the computer side, I am still learning a lot about the technical details every day, and I am happy to do it on this conference.