Extracted from: Archiving 2014: May 13-16, 2014, Berlin, Germany; final program and proceedings, Springfield (Virginia) 2014, pp. 15-19.

Crowd-based appraisal and description of archival records at the State Archives Baden-Württemberg

Kai Naumann; Landesarchiv Baden-Württemberg – State Archives Ludwigsburg; Germany
Franz-Josef Ziwes; Landesarchiv Baden-Württemberg – State Archives Sigmaringen; Germany

Abstract

Appraisal and description are core processes at historical archives. This article gives an account of innovative methodologies in this field using crowd-sourced information to (1st) identify which files are of interest for the public, (2nd) enable agency staff to extract and transfer exactly those files selected for permanent retention and (3rd) ease the description and cataloguing of the transferred objects. It defines the extent of outsourcing used at the State Archives (Landesarchiv Baden-Württemberg LABW), describes case studies and touches issues of change management. Data sources are government databases and geodatabases, commercial data on court decisions, the name tags of German Wikipedia, and bio-bibliographical metadata of the State Libraries and the German National Library.

1. Introduction and overview

Archives in Baden-Württemberg, and in Germany in general, are a stronghold of democracy and rule of law by securing permanent access to selected government and private records. Our records open authentic perspectives on good, bad or just usual memories of the past. The Landesarchiv Baden-Württemberg (LABW) has about 190 employees at seven locations (mostly called State Archives) throughout Baden-Württemberg. Whenever agencies want to delete records (for privacy reasons or because they are obsolete), archives are authorised by law (Archives Act of 1987) to select which records will be stored forever. The LABW aims to document government and administration business, but also society as a whole.

Due to limited storage capacities, the LABW can only retain between one and two percent of all records produced by public agencies and courts; that is 1,000-2,000 shelf metres of paper files, or up to 400.000 individual folders per year. Usually, records are transferred between 10 and 30 years after the last piece of paper entered the file. Electronic records have also started flowing into the archives, to sooner or later become a torrent.

In the past, neither archivists nor agencies have had enough means to look at every single file and select the "historic" subjects. Archives and agencies agreed to retain blocks of files ranged in agency shelves. E.g. the files marked "A123", pertaining to senior careers, were selected, while other files on human resources marked "A121" through "A129" where to be shredded. With personnel records, we resorted to requesting only records where family names started with D, O or T. Another example are courthouse registries, where only the files of the years 1980, 1985, 1990, and so on were chosen to be preserved. Of course, archivists never failed to ask agency staff for additional nominations of famous cases, but mostly in vain [1].

The only known remedy was what we call autopsy, i.e. the dusty and tedious task of personally flipping through files in attics or basements. Sadly, this only worked for a tiny portion of all files marked for deletion, due to lack of staff. In the subsequent stage of ordering and description of the transferred records, the archivists were on their own for the keying of metadata and the production of finding aids.

This article shows how these robust but imprecise methods have been partially overcome at our institution and what can be done to use our approach on a larger scale. It starts with a short description of the approach, continues with its history, explains three case studies and concludes with thoughts on what we learned and suggestions on how the shift towards crowd-based appraisal and description can be established easily.

2. What are we sourcing out? Who are our crowds?

The notion of crowdsourcing used in this paper is slightly different from the mainstream one where crowdsourcing means transferring tasks to the crowd, i.e. volunteers on the web. Its parent term outsourcing sometimes has a bad name in the business world.

The processes of selecting the historical records, called appraisal, and describing them in finding aids are core processes at archives. At first sight, the outsourcing of this job to a crowd seems like self-disempowerment. Curiously, the methods the LABW has explored rather confirmed than challenged our position as appraisal experts. This is largely because we decided to use crowdsourced material only in addition to the established ways. It is and will be limited to record types that are suited for this method, in having adequate metadata on every single record in a database. As an output of crowd-based appraisal, records lists are generated and handed to the archivist in charge, who can add records to the list and also withdraw records.

The crowd knowledge used has mostly not been generated for our purposes, but we request it from external, independent groups who register information on events, objects, or persons. Those groups can be as wide as world society, like Wikipedia, or limited to a tiny number of agency staff (open vs. closed crowds). The following chapters present case studies in which the sources will be presented in detail.

3. How to use the collective wisdom of Wikipedia for appraisal

For historians, court decisions and associated records are an indispensable source on conflicts in a past society. For a longer period than in other record types, the LABW has stored nearly every file created. This became impossible with the records produced from 1950 onwards. The method shifted to retention of certain file groups. Appraisal models were created which privileged “notorious justice”, i.e. murder or political crime, against everyday causes like divorces, theft, or libel. This worked because courthouse registries stored “notorious justice” records separately from their ordinary work. But it was a well-known problem that “everyday” and “next door” justice often escaped us into the shredder. There were also little means to spot cases of this kind in which notorious people had been involved.

Ways to overcome this were first spotted in the mid-nineties, when courts were introducing workflow management databases [2]. Using the metadata in those databases to pick suitable files sounded simple. In reality, it took over fifteen years to overcome several obstacles: concerns over data protection, budget limitations and a lack of persons who were able to work with database records (which have little in common with paper records). It was then agreed to only send us data on closed cases in order to prevent the betrayal of secrets.

Surprisingly, after the first database transfers from the courts had been distributed to archivists, there was little response. The amount of records was overwhelming, yet there were no tools to filter and sort the data, and documentation on the data was scarce.

How do you find criteria and tools for database appraisal? The first challenge was to efficiently apply filters on the metadata supplied. The State Archives of Saxony, in the east of Germany, was the first to respond to this demand by producing and sharing a tool for those metadata, called jBewerter (“jAppraiser”). It was able to select some keywords (like “electricity theft” or “poaching”) and to display certain age groups or to concentrate on lawsuits terminated by formal decisions [3]. Our staff at the LABW was happy about it, but also demanded more freedom in defining their own criteria.

The second challenge was to select the cases the public was interested in. How do you define those cases and collect metadata on them? Our user statistics showed that records on persons were the most popular ones; therefore, metadata of this kind were our first goal. The staff took blind alleys on their way like a project that was about encouraging the public to anonymously suggest persons and file references on an on-line form, which was specifically intended for this purpose. It never came to life.

The best solution was much easier to achieve: in 2009/10, Franz-Josef Ziwes had to appraise an enormous amount of 27,428 personnel files from a large agency. The agency had kindly invested a lot of time in preparing a spreadsheet with names and dates of birth of all employees. In the first stage, Franz-Josef invested his own knowledge on persons and asked historians for help to examine the list. This provided files on 148 persons. But as he felt that database methods might enhance the result, he finally acquired database sources of which the German language version of Wikipedia was the most successful one. The Wikipedia community offered a way to acquire a dataset of first and last names, short descriptions, and dates of birth for hundreds of thousands of people, which other people had chosen to key into the online encyclopedia. On the 27,428 files mentioned above, it supplied another 113 hits which had not yet been spotted in the first stage.

The numbers showed that the massive collective wisdom of Wikipedia authors has a great advantage over the expert knowledge of historians and archivists. Another less obvious, but even bigger advantage of using Wikipedia data is the democratic legitimation of our appraisal. Everybody has the right to create a Wikipedia article on a person, and the Creative Commons licence guarantees free access and reuseability of those data [4].

In 2012, the LABW started to accumulate the Wikipedia data by regularly harvesting the personal data and continuously adding them to its own list, without taking into account the continuous deletions on productive Wikipedia. Alongside the global crowd of Wikipedia, we also acquired the person-related data of the State bibliographical reference system (Landesbibliographie) maintained by the State Libraries and the Statistical Office. The whole data collection has been labelled “Database on person-related appraisal” (DpA, or DpB in German).

3. The Notion Potion case study

During 2012 and 2013, several smaller projects were started at the LABW to do database-appraisal on metadata of various courts and other person-related records with encouraging results. In early 2013, Kai Naumann acquired another metadata source from JURIS, the German market-leader on court decision metadata. The company supplied the data, reaching back as far as 1946, for use at all German state archives for a small fee. The data

Selector name	**Description**	**Hard/ soft**	**Concentration**	**Proportion**
Mass offences	OFFENCE contains “damage”, “theft”, “deceit”, …	soft	0,3 %	9 %
Non-rare offences	OFFENCE contains “forgery”, “copyright”, “libel”, …	soft	1 %	25 %
Rare offences	OFFENCE contains “shares act”, “animal protection”, “prison mutiny”, …	soft	3 %	20 %
Violent felonies	OFFENCE contains “murder”, “breach of the peace”, “terror”, …	hard	100 %	7 %
Young offenders	DATE_BIRTH lesser than actual year minus 16	soft	10 %	2,5 %
JURIS hits	FILE_REFERENCE matches references in JURIS data	hard	100 %	< 0,5%
DpB hits	NAME, FIRST_NAME and DATE_BIRTH match DpB data (not only offenders, but also victims)	hard	100 %	9 %
Long term trials	More than 8 years between DATE_REGISTRATION and DATE_DECISION	soft	30 %	0,5 %
…	…	…	…	…

Figure 1. Some sample selectors of those used in the Notion Potion project on criminal court data. The proportion values are related to the total of records returned and vary depending on which kind of courthouse data is processed.

can be used to select all records whose headnotes and provisions had been found valuable enough by judges and barristers to have them added to the JURIS database.

Afterwards, a larger team formed in summer 2013 to apply all lessons learned on the appraisal of criminal case records, which continuously kept archivists busy. The project's name Wunschpunsch (Notion Potion) was lent from a children's book by the German author Michael Ende and fits well because it really involved something like a recipe.

The Notion Potion was implemented in an office database format that was SQL-enabled. We agreed to set up several queries on the data expressing various aims of appraisal, called selectors or ingredients (Figure 1). The selectors only return data for records whose retention period has ended based on a date variable the user has to supply. Today, sixteen selectors have been defined. Most of them only make use of the vocabulary provided by the courts, i.e. the free-text terms for offences, dates, or fixed reference numbers for statistics. Others make use of DpB and JURIS data.

In a second stage, we counted how many records the different selectors returned and started to balance all the ingredients in a sensible way. Some selectors had to be reduced by narrowing the criteria or by randomly selecting only part of the selected records with a "concentration" value. The random selection is carried out by filtering out sequential reference numbers ending with "1" or "11". We distinguished "hard" selectors that should not to be thinned out from "soft" ingredients on which we thought reduction was harmless. The proportion values shown in the table are related to the total of records returned and vary depending on which kind of courthouse data is processed.

We adjusted the selectors to retain a subset of two percent of all paper records for permanent retention, but that quota can easily be adapted to larger or smaller proportions, depending on local archival policies. Since some records were chosen by more than one selector, we also made a de-duplicated total of all file references requested.

The Notion Potion recipe that resulted will be revised from time to time. It is part of an office database file which can be connected to the local courthouse datasets and applied to return spreadsheets which archivists can send to the courthouse or use in autopsy.

It was an exciting experience for all participating colleagues to shape the selectors and look at the datasets they returned. Many were surprised that their – already large – perspective on the court's work actually broadened once again. It motivated others to acquire more skills on database software. The Notion Potion has also strengthened their position towards the courts, because they can now exactly define which and how many records they want to have transferred.

4. The Notion Potion as part of a bigger appraisal concept

Users from the social sciences might feel uneasy about the way in which mass phenomena are under-represented in records selected by Notion Potion. In fact, the LABW has acknowledged this gap and has decided to retain not only the few chosen paper records but also the whole of the courthouse database transfer containing data about offenders, victims, and barristers. But on the paper record level, it was decided that the archives can only achieve a diverse representation of crime economically by over-representing the extraordinary.

These decisions are based on the LABW's concept on appraisal for person-related records. In 2008, our appraisal officers agreed on a policy for that kind of records based one five use cases [5]:

1. Use a total of all records produced
2. Use a statistically analysable subset
3. Use a set of records on average/typical cases
4. Use a set of records on outstanding cases
5. Use a set of records in order to see how the agency worked (Schellenberg's "evidential value")

Applying this to criminal case records, we will mainly facilitate use case 1 with database contents and use cases 3 to 5 with the few surviving paper records. Use case 2 does not apply because no subset is necessary, given the totality of datasets.

5. Crowd-based description: leveraging archival description trough crowd-maintained name authority files

Having been transferred from the agencies and courts into the state archives, the appraised records need to be described and made available for research. The DpB is useful for this task, too. For several years, the LABW has established descriptors as an

```
#FORMAT: BEACON
#PREFIX: http://d-nb.info/gnd/
#TARGET: http://www.landesarchiv-bw.de/plink/?gnd={ID}
#VERSION: 0.1
#FEED: http://www.landesarchiv-bw.de/beacon
#INSTITUTION: Landesarchiv Baden-Württemberg
#CONTACT: Email: landesarchiv@la-bw.de
#MESSAGE: Online-Findmittelsystem, Landesarchiv Baden-Württemberg
#TIMESTAMP: 2014/03/02
100001718
100006000
100011071
100019552
100021549
100027385
100029752
10003148X
```

Figure 2. First lines of a BEACON file at Landesarchiv Baden-Württemberg (LABW).

additional procedure in the cataloguing of records. This provides not only an index of persons and an index of places, but also a specific method to link up archival information systems with a semantic web developed by the big libraries. We tag the record description with the ID number of the German Name Authority File (Personennamendatei, PND) [6]. Since April 2012, this PND is part of the Integrated Authority File (Gemeinsame Normdatei or GND).

The PND is an authority file of people, which primarily serves to access literature in libraries. The PND was built up between 1995 and 1998 and is maintained by the German National Library. For every significant person, a record can be created with name, birth and occupation, receiving a unique identifier, the 9 or 10 digits PND Number. As a standard file for persons or personal names, the PND enables a uniform bibliographic record of literature in libraries and should enable more efficient search for records on people in the holdings of libraries, archives and museums in the entire German-speaking world. Of the total 7.1 million PND records, 2.6 million refer to an individual; the others are merely standardised catalogue name terms.

The two sources of the DpB mostly contain the PND Number of a person. By a data comparison with full name and date of birth, the PND number belonging to a significant person described in the finding aid can be fairly easily determined. The archivist only has to add the respective PND number to the data record of his or her description and export it to the Internet. All you need now is a so-called BEACON file. This simple textual file format has been developed to help websites signal that they provide content to normative data and at the same time show the path to these data (see Figure 2). With a PND BEACON all websites providing

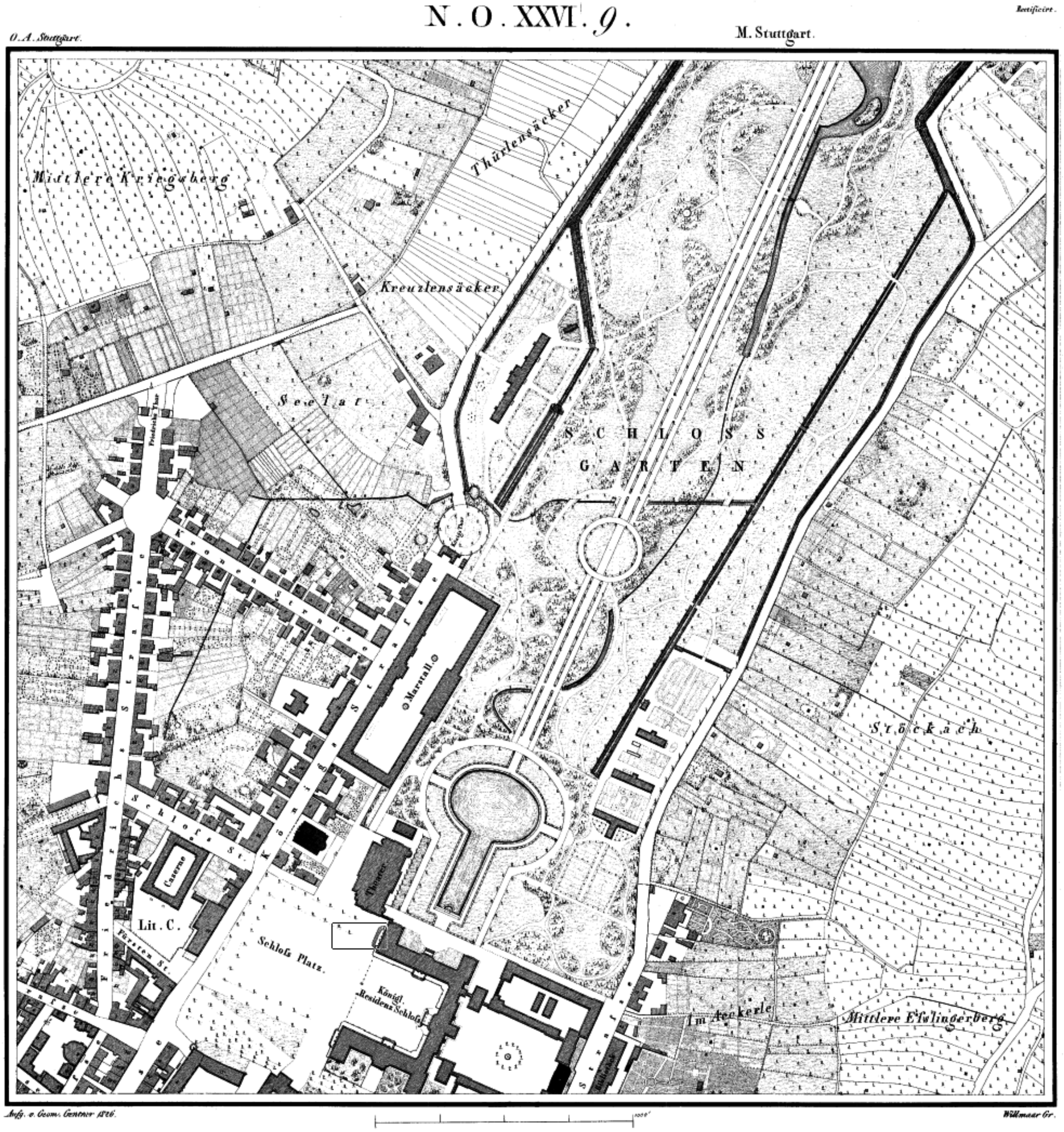

Figure 3: One of 16,687 map images auto-described by modern-times surveying geodata. This map shows the "Schlossgarten" park at Stuttgart, now a densely settled city centre, in 1822.

content to normative data can be linked to each other dynamically. The principle is rather simple. Any cultural institution which provides normative data on the web generates a simple text file with one table column. The header contains some metadata such as format, URL resolution of the decided target, contact details etc. Then only the individual PND numbers follow. The file enables a 1:1 or 1:n mapping from persons to associated records metadata.

At this time, more than 150 websites successfully use this lighthouse principle. The most prominent and widely used page is probably the “Wikipedia Personensuche” (Wikipedia People Search), developed by Christian Thiele and available on http://toolserver.org/~apper/pd/about.php. This webpage is linked with any biographical article in Wikipedia and is hosted on the Wikimedia Tool server. Every website which is registered on the Wikipedia page http://de.wikipedia.org/wiki/Wikipedia:BEACON as BEACON provider is included in this search engine and becomes part of the hit list. By a concordance to Virtual International Authority File (VIAF), the beacon resolver of the PND is even linked with other international standard files, such as the Library of Congress Name Authority File (LCNAF).

The LABW could not let pass this unique opportunity of free publicity, and in July 2012, it posted the first beacon file with slightly more than 4200 PND numbers. Today, there are about 10700 PND numbers, and the trend continues to rise. In the future, all archives might dynamically connect their websites to each other in this way by embedding freely available BEACON resolvers into their own websites, e.g. by the SeeAlso-service on http://beacon.findbuch.de.

6. Using closed-crowd geodata on map description

The approach of using other people’s data has also been used on description itself, saving four years and eight months of work. It was implemented with kind permission of the State Surveying Agency (Landesamt für Geoinformation und Landentwicklung LGL). A large portion of 16,000 maps was to be described.

The maps, drawn in the 19th century for cadastral purposes, where laid out in a grid which was available as machine-readable vector geodata, having the map reference numbers as metadata. On the other side, there were datasets for names of towns, villages, homesteads and even land parcels produced by local surveying authorities. This wealth of over 80,000 place-names could be associated with the map reference numbers, thereby avoiding a description by visual inspection, which would have consumed fifteen times more time (five years) than the method applied (four months).

7. Reflections and outlook

The way into crowd-sourcing presented in this paper happened away from the mainstream. It was not conceived at headquarters and deployed locally, but it rather grew bottom-up when concrete issues arose on the operative level. In order to be successful, our previous database experience was essential.

The Notion Potion pilot scheme has a broader potential [7]. Institutions which want to copy the Notion Potion method should acquire basic technical knowledge, develop a test bed, then gather all persons concerned and collectively improve it until it is ready for use. Senior managers who want to create preconditions for this or similar developments should create a climate of trust with agencies, which enables the transfer of metadata and foster database classes in education and on-the-job training.

Reflecting on recent archival theory of appraisal, the Notion Potion seems like a mixture of the known methods. It has aspects of sampling since it involves random selections. It is partly macro-appraisal, because it relies on generalised, functional criteria like big crime vs. trivial offences. It is also micro-appraisal, because each record is appraised separately, though not by a human.

The LABW will continue to use crowdsourcing in the described ways. There are plans to set up a database environment on an intranet server to facilitate the use of DpB, descriptor generation and Notion Potion. The LABW will soon also become operative in the field of classic crowdsourcing, i.e. projects open to everybody on the web. The first project will be about name-tagging lists of World War II victims in Baden-Württemberg that were machine-typed between 1946 and 1982.

References

[1] R. Kretzschmar, Archival appraisal in Germany: a decade of theory, strategies, and practices, Archival Science, 5 (2005) 2-4 pp. 219-223.

[2] U. Schäfer, Büroautomation in der Landesverwaltung Baden-Württemberg – Strategisches und operatives archivarisches Handeln am Beispiel der Justiz, Arbeitskreis Archivierung von Unterlagen aus digitalen Systemen, Münster (Germany) 1997, pp. 31-48, esp. 34-39. http://www.staatsarchiv.sg.ch/home/auds/01.html.

[3] B. Nolte, Effiziente Überlieferungsbildung durch Nutzung der Anwendung “J-Bewerter” für Strafverfahrensakten. Erfahrungen des Sächsischen Landesarchivs; E. Koch: Welche Morde und wieviel Diebstahl braucht die Zukunft? Überlegungen über das Archivieren von Strafakten im Zeitalter neuer datenbanktechnischer Möglichkeiten, Proc. of Workshop „Ziele und Methoden archivischer Bewertung“ (Aims and methods of archival appraisal), Stuttgart, Dec. 1st 2010, http://www.landesarchiv-bw.de/web/52498.

[4] F.-J. Ziwes, Wikipedia und Co. statt Sisyphus?: Konventionelle und digitale Hilfsmittel zur qualitativen Bewertung von Personalakten, Archivar, 63 (2010) 2 pp. 175-178. For practical guidance, follow F.-J. Ziwes: Überlieferungsbildung und die Intelligenz im Web. Digitale Hilfsmittel bei der Bewertung personenbezogener Unterlagen, Proc. of Workshop (cf. ref. 2).

[5] A. Ernst/C. Keitel/E. Koch/C. Rehm/J. Treffeisen, Überlieferungsbildung bei personenbezogenen Unterlagen, Archivar 61 (2008) 3 , pp. 275-278.

[6] Deutsche Nationalbibliothek. Bibliografische Dienste. Stand 1. Mai 2012, Frankfurt, Leipzig 2012, p. 33.

[7] C. Ferle, Ein integriertes, digitales Bewertungsmodell am Beispiel eines Vorgangsbearbeitungssystems, Presentation at AUdS Workshop, Weimar, 11th March 2014, http://www.staatsarchiv.sg.ch/home/auds/18.html.

Author Biography

Dr. Kai Naumann is senior archivist at the Ludwigsburg State Archives, a division of the Landesarchiv Baden-Württemberg (LABW). His responsibility is archiving of born-digital records. He is member of committee of the German State Archives Conference.

Dr. Franz-Josef Ziwes is deputy director of the Sigmaringen State Archives, a division of the Landesarchiv Baden-Württemberg (LABW). For many years, he has planned and implemented computerised ways of facilitating archival business processes.