

Keeping Electronic Records Accessible - How The Federal Archives Of Germany Preserves The Digital Heritage Of The German Democratic Republic: A Case Study

Andrea Hänger

About 200 datasets (with some 200,000 files) stored in the digital repository of the Federal Archives of Germany were created in the German Democratic Republic between 1970 and 1990. Computerised databases covering the general areas of statistics, economics, agriculture, education, penal registration, and labour have all been preserved. These records are frequently used by scientific users as well as by agencies and individuals to verify legal claims. This paper seeks to present the acquisition, preservation and accessibility of these records. It discusses the question of authenticity and assesses the prospects for presenting and preserving the data with modern exchange standards like the Metadata Encoding and Transmission Standard (METS).

Acquisition

Following formal unification in October 1990, East German government agencies and institutions that were not taken over by federal agencies or by one of the newly established Länder were either privatised or dissolved. As a result, many of the State data-processing centres were shut down, and in the dismantling process, data holdings were often systematically destroyed or relocated within new private companies. The Federal Archives immediately set up an active acquisition policy and succeeded in rescuing a significant number of historical and evidential records. Above all, it was important not only to acquire data, but also to safeguard documentary information that allowed this data to be read and interpreted. The acquisition was, and still is, an ongoing process. Over the years hundreds of magnetic tapes were found in attics or in abandoned offices. Even today, floppy disks are found in paper records in the course of description. The latest discovery is a dataset containing the findings of an investigation committee of the last GDR-parliament, the Volkskammer. This committee was established to throw light on accusations of corruption against the government. Data was stored on several 5.25" floppy disks in the format Redabas, the socialist version of dBase. One of the most important datasets is the "Central Executives (Cadres) repository" which holds data

gathered during the 1980s in the GDR's Council of Ministers. With records covering approximately 700,000 individuals including the whole 'functional' elite and the upper echelons of the GDR's service class (but excluding full time party functionaries and officers in the military and the security apparatus) this unique data source provides us with full records of their social and political family backgrounds, their family situation in the 1980s, their occupational careers, educational backgrounds, party and organisation affiliations, status in the nomenclature and further information regarding their position as cadres (such as foreign language skills, entitlement to travel into non-socialist countries etc.). Other examples of important datasets are the "1971 and 1981 Census dataset" or the "Petition dataset" containing the description of more than one million petitions addressed to the GDR-government between 1979 and 1989.

Preservation

Following the first acquisitions, a strategy for the long term preservation had to be defined. According to Margret Hedstrom preservation means: "Retaining the ability to display, retrieve, manipulate and use the digital information in the face of constantly changing technology"¹. For archival purposes it is not only preservation for the months and years to come but it is also long term preservation. Long term means that it is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long term may extend indefinitely.

In general there are two different strategies for long-term preservation of born digital records: migration and emulation. Born digital records are those whose meaning or usability arise from and rely on their being encoded in digital form. They cannot be stored on anything other than digital media without losing their whole functionality. Migration can be broadly defined as the transfer of digital information from one hardware/software platform to another, or from one generation of digital technology to another. Migration is almost always an ongoing process. Rarely can one migration result in permanence of data; instead, migrations must take place periodically in order to ensure that information remains "evergreen."

¹ Margret Hedstrom, Preserving digital information, in: Long Term Preservation of Electronic Materials A JISC/BRITISH LIBRARY Workshop as part of the Electronic Libraries Programme (eLib), <http://www.ukoln.ac.uk/services/papers/bl/rdr6238/paper.html>

Emulation as the second alternative attempts to preserve the original software or hardware environment. Ideally, emulators reproduce the “look and feel” of the original electronic document by mimicking the software or hardware with which the document was produced. The user can thus access the document in its original format, with all functionality preserved. In the perspective of archival sciences this is without any doubt the best way because it keeps as much of the original as possible. However, it is also the most complicated and expensive way. The choice of the strategy depends on the degree of functionality which needs to be preserved and on the resources an archive can invest in digital preservation.

It would be fair to say that, as yet, no single strategy has emerged as a clear cut solution, and digital preservation remains a technically complex and resource intensive process. The first aim should be to keep data open to any questions which could be raised over the course of time. In the long term new communities of users will emerge with needs and expectations that differ from those of the communities that created the digital content.

Both migration and emulation represent ongoing preservation commitments – an archive must be prepared to provide resources on a cyclical basis in order to maintain the accessibility of its records. Even ignoring the issues of preserving content in a meaningful manner, the physical storage media are themselves subject to obsolescence and deterioration, and must be regularly refreshed.

The Federal Archives has decided to follow the migration strategy for the media and for the formats. For the GDR-datasets ASCII was chosen as preservation format. ASCII is undisputed the safest way to keep data, but only the bit stream without any information about former functionality and form is preserved. This information has to be kept in supplementary documentation.

In many cases data structures had to be reconstructed. A lot of research had to be done to identify structures and codes. In order to save storage space which was extremely expensive when these records were created the information had been compressed as much as possible. Mostly data was encrypted and packed, so that the conversion into an archival format could not be managed by commercial conversion programs. Each dataset demanded special programming. In some cases the software engineers who had written the algorithms had to be found because they were the only ones who could help with the encryption. In 2003 the Federal Archives developed a special conversion tool which migrates the original file from EBCDIC to

ASCII and XML. It has an unpack routine for packed data and converts hexadecimal into binary values. With this tool a simple analysis of the values of each field is possible, so that file structures and codes can be validated. The most important function consists of the non-detachable connection of the data, the description of the structure and the codes. This function is able to compensate for the disadvantage of ASCII that only the bitstream is preserved.

During the processing of the datasets, often anomalies were identified in the data. Information about content validation and possible constraints on the reliability of data are documented in the finding aids. Sample checks were carried out to compare the transformed data against the original data. These included carrying out plausibility checks on the data, comparing the value of specific fields and checking that the overall number of records and fields remained the same.

Accessibility

The GDR-datasets are frequently used not only for scientific but also for evidential purposes. Thus, the conversion tool is also able to replace codes. That is to say that the numeric codes are replaced by plain text, for example "German Democratic Republic" instead of "111". This is especially important for datasets which are frequently used for official and individual purposes and which contain highly sensitive personal data. Whereas for example the Social Courts are provided with the "GDR Company Register dataset" as a whole to carry out the retrieval themselves, it is, as a matter of course, not possible to deliver copies of datasets containing personal data. For enquiries related to these datasets the research can only be carried out by the Federal Archives.

The highest number of enquiries is related to the "Detainee and Prisoner dataset". The tables include information about the family situation, the term of imprisonment, the punishable act, and every event (including illnesses) which occurred during the term of imprisonment. Much time and effort was spent on the conversion and documentation of this dataset. Many codes have been replaced by plain text. The advantage of plain text spreadsheets over encoded data is that no expert is needed to carry out a research.

To support fast access another tool was developed which adapts the structures of files with the same content but varying structures. For example the "Detainee and Prisoner Dataset" contains annual datasets from 1980 up to 1990. The structure has

evolved over the years. For research purposes the tool normalises the structure and reads the files in a mySQL-database. As a consequence, an enquiry can be answered by a single retrieval. This tool is also able to compare the different files and to identify those which are identical.

For scientific users anonymised copies are produced. The conversion tool is able to extract the fields to be saved and anonymise those fields that contain sensitive information. Most of the users work with copies of data delivered on CD-R. These users are participants of bigger research programmes where data is processed in modern databases. They only come to the Federal Archives to consult the supplementary documentation. Only few users use the on-site public access of data. The dataset description catalogue is presented on the website of the Federal Archives. A web-based public presentation system is not yet planned because access conditions for many datasets specifically prohibit free access.

Authenticity and reliability

The question how to keep and prove the authenticity of electronic records is a core issue not only for archivists but for the information society as a whole. As a consequence one can observe that electronic records management programs have to face security requirements which clearly exceed those for paper records.

In turn, the experience with the GDR-data collection is completely different. The records are frequently used for evidential purposes. Citizens of the former GDR use the records to establish their claim for refund, compensation or entitlement for a pension. To give you some examples: One of the most important datasets is the dataset "Corporate Working Capacity" which contains the individual-level information for a high percentage of the former GDR workforce, including details on education, training and employment for about 7,25 million individuals. This source often serves as proof of periods of employment.

As mentioned above, the highest number of enquiries is related to the "Detainee and Prisoner dataset". Information especially about diseases suffered during the term of imprisonment can help former prisoners to get compensation. It exists in electronic form only and it is the only source a person concerned can rely on when applying for compensation. In the first instance local authorities or regional government departments decide about the application. If it is defeated the applicant has the possibility to challenge the decision and to make an application for judicial review.

As shown above the acquisition of these datasets did not comply with standards and guidelines about secure data transfer. But no authority or judge has, to today, doubted the authenticity and reliability of the records. Apparently the careful documentation about the context, the validation, transformation and processing of the data is adequate for evidential purposes.

This experience confirms the policy to avoid short- or mid-term remedies like digital signatures for long-term issues and to give priority to organisational rather than technical provisions to keep the authenticity. The Federal Archives is taking an active part in a multidisciplinary initiative which is modelled on the RLG's "Digital Repositories Certification". The goal of this project is to set up requirements for digital repositories which are able to reliably store, migrate, and provide access to digital collections.²

METS

It is a prerequisite for keeping reliable and authentic data that the accompanying documentation is complete. As shown above, to start with, the Federal Archives developed a tool which non-detachably connects the data, the description of the structure and the codes thus avoiding that the link between these different sources gets lost. If one of the three parts is missing the data cannot be interpreted anymore. But from an archival perspective it is also a core requirement that context and technical information are preserved with the data to make sure that at any time the whole history of the file can be reconstructed. This includes information about the administrative context of the creating agency, the way the data was originally captured and validated and the documentation of the processing of data in the archive, for example about content validation, sample checks, migration, refreshment etc. Today this information is kept separate from the data. The Federal Archives is currently investigating if the Metadata Encoding and Transmission Standard (METS) could be a way to bring data, context and technical documentation together. METS is an initiative of the Digital Library Federation. It provides an XML document format for encoding metadata necessary for both management of digital objects within a repository and exchange of such objects between repositories and their users.

² nestor: Network of Expertise in Long-Term Storage of Digital Resources:
<http://www.langzeitarchivierung.de/index.php?newlang=eng>

Originally intended for presentation purposes³, METS is increasingly used for preservation purposes. A METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model. It uses XML as a flexible, open standard with widespread support, combined with the ability to separate content from presentation in a manner which offers many advantages to the archivist.

A METS document consists of seven major sections: 1. the METS-Header describing the METS document itself, 2. the Descriptive Metadata, 3. the Administrative Metadata, 4. the File Section listing all files which can be referenced or contained within the METS document (in XML or as Base64 Binary) , 5. the Structural Map which describes the hierarchical structure of the digital object, and links the elements of that structure to content files and metadata that pertain to each element, 6. Structural Links and 7. Behaviour.

The advantage of METS is that the descriptive and the technical/administrative metadata are not defined internally, but they point to external standard schemas. These could be international ones like Dublin Core or EAD, but also national standards for records originating from electronic records management systems. Also the technical metadata can be represented by international standards like PREMIS.⁴ PREMIS for example offers the possibility to register every "event" that occurs while transforming, processing or describing data. METS offers the possibility to link together the descriptive and the technical metadata and provides for a coherent documentation not only of the content and context but also of the preservation process.

Lessons learned

The experience with the GDR-data can be seen as a case study which shows what happens when the cooperation between the data producer und the archive fails. To start working only when records are given up without a standardised transfer is a difficult and expensive task. In the years to come we anticipate that the majority of our accessions will originate from electronic records management systems, and will

³ In the research project "Digitized Archives in Online Finding Aids" supported by the Andrew W. Mellon-Foundation a pilot application for the presentation with METS is currently developed at the Federal Archives: www.daofind.de.

⁴ PREMIS (Preservation Metadata: Implementation Strategies): <http://www.oclc.org/research/projects/pmwg/>

arrive with standardised metadata and comprehensive audit trails, but the lessons learned by the GDR-experience will need to be kept in mind:

Archivists have as their mission the preservation of records of continuing value and the provision of access to those records. However, that does not imply that archivists should remain on the sidelines until the time comes when an organisation feels that it no longer needs its records. Representative and relevant archives are based on records that are created and managed well by the creating organisation. This axiom is of greater importance in the electronic era where lack of planning can doom electronic records to an early grave. Archivists must be involved early in the life cycle of records if they are to have an impact. If action is not taken, there is a significant risk that society will lose generations of historical and evidential records as archives become increasingly impoverished. Archives rely heavily on data producers to provide complete and accurate documentation when they deposit data and to comply with other requirements, such as file structures and formats, transfer media, and requirements for protection of privacy and confidentiality. The entire enterprise of digital archiving assumes some degree of cooperation between producers of digital information and the archives. When data producers do not comply with submission guidelines, archives incur additional costs in preparing the data for preservation and dissemination, They experience delays between ingest and release, and assume risks if data that does not meet quality assurance standards is released.

The example of the GDR-data underlines that archives wishing to set up a strategic vision for electronic records must take account of two perspectives: There is the cultural side of their role, which focuses on access, learning and being culturally inclusive. But they have also to be aware of their evidential role, i.e. their potential to provide evidence of rights of democratic access

For both, the cultural and the evidential role authenticity is a major issue. METS could present – besides other advantages – one possibility to safeguard authentic records over the long term.