THE DIGITAL ARCHIVES OF THE GERMAN FEDERAL ARCHIVES FROM "MACHINE-READABLE DATA" TO "BORN-DIGITAL" RECORDS

Dr. Sina Westphal
German Federal Archives

1. Introduction and History of the Digital Archives

The German Federal Archives (approximately 700 staff members) is the National Archives of Germany. Founded in 1952 it has seven departments in nine locations. It preserves historically significant federal archival documents going back to the year 1411, whereas the majority of the archives has been created since 1867.

In the following lecture I will address the question, which events in recent German history have affected the development of the Digital Archives and which specific organizational and technical solutions arose as a result of these particular circumstances.

The Archives has long since been in charge of digital records: The legal foundation was established in 1988, when the "Law on the Preservation and Use of Federal Archival Documents" was enacted. It states that the Federal Archives has the legal responsibility of permanently preserving the so-called "machine-readable data".

After the German reunification (1989/1990) the Archives received electronic materials, originating from the liquidated government agencies of the former German Democratic Republic (GDR) of the 1970'ies and 1980'ies.[1] In the former GDR "machine-readable data" from government agencies was often stored in centralised state-owned data-processing centres, many of which were closed after the reunification.[2] Thus the stored content was in danger of being destroyed or getting lost to private companies. The Federal Archives in unison with the Federal Government had to act. Thus in 1990 there was an amendment to the "Law on the Preservation and Use of Federal Archival Documents" which entitled the Federal Archives to acquire records of the former GDR. Hence within the following ten years circa 180 databases (Fachverfahren) were transferred to the Federal Archives. Among this data was for example a database, which contains citizens' petitions, directed at the State Council of the GDR (1979-1989) or data from the commission of the people's Chamber (i.e. the unicameral legislature of the GDR) concerning corruption and abuse of office. Large amounts of the transferred data were created by the Central State Administration for Statistics.[3] It was mostly encoded in EBCDIC (not in ASCII), which was a format typically used on (mainly IBM) mainframe databases from the 1950'ies to the 1980'ies. Thus in 1991 a young archivist of the German Federal Archives was commissioned with the preservation of this data. Moreover he was assigned to acquire new digital archival materials from the federal agencies. This step enabled the Federal Archives to gain first-hand experiences in a field, which was still in its infancy.[4] During the following ten years he worked hard – in league with three colleagues who joined him in the early 1990'ies – to preserve the new data and to get all the background information he could. Not all the information, which was needed to read the in EBCDIC encoded data, had found its way to the Federal Archives. That's why the team made an effort to get all the missing metadata and code lists: For this purpose they tracked down former employees of the above mentioned data-processing centres and convinced them to help the Federal Archives. But they had more basic needs too, which had to be addressed. At the beginning there was for example no suitable hardware and it took some time to get all the things, which were required to be fully operative. In 2001 some organizational transformations within the Federal Archives led to a change of responsibilities. Thus the task of preserving the digital archival materials was assigned to another unit.

Five years later three issues led to the constitution of the 'Digital Archives' as a special unit of the Federal Archives: Firstly the aim to lastingly preserve the "machine-readable data" from the 1970'ies to the 1990'ies. Secondly, the aim of maintaining the electronic records, that were soon expected then. Thirdly, the publication (2003) of the Open Archival Information System (OAIS) Reference Model as a(n) (ISO) standard and the Preservation Metadata

Implementation Strategies (PREMIS) as well as some other international developments within this branch of the archival sciences.[5]

At that time, the challenge was to create a business model for digital preservation and custody of these different archival materials, which has been established in 2008.[6] In the following I will describe the organizational and technical structures we have built so far.

2. Organizational and technical solutions

2.1. Mission, aim and services offered by the Digital Archives

Our core mission remains to make the digital records of the German Federal Archives available to the public. In this way we intend to encourage the dissemination of knowledge and understanding of history as well as to enhance transparency by providing the public with information about the government's dealings. Thus our main aim is the preservation of our heritage.

At the present time we are able to render most of the services described in the OAIS Reference Model (see below the description of the Digital Archives' architecture). In the following I will dwell on the services the Federal Archives offer, referring to the functional model of the OAIS.

Unlike some other National Archives we do not offer training courses in records management to the staff of government departments,[7] but organize conferences on relevant topics and provide advisory services regarding both the management of electronic records in general and the implementation[8] of specific Electronic Document and Records Management Systems (EDRMS).

At present staff members of the Digital Archives have the authority to appraise specific archival collections independently. In future the Digital Archives will procure mainly technical and organizational support during the appraisal process, while the unit within the Federal Archives, which is responsible for the respective archival collection, appraises the records. Currently we structure the captured records originating from shared folder file systems and enhance it with metadata[9] previous to the ingest, as will be explained in some more detail below. Later the staff of the government agencies, to which the transferred records belonged, might do this job and afterwards send the finished SIPs to the Digital Archives. Electronic records from EDRMS are imported into the Digital Archives without pre-ingest by an automated transfer process.

At this point in time the archivists have access to the archival information packages (AIPs) within the system, while there is still no defined procedure transforming born-digital AIPs into dissemination information packages (DIPs) within the framework of a user access system. Likely we will hold DIPs in our digital repository and regenerate the information packages according to requirements. In this case all SIPs will be automatically transformed into AIPs and DIPs during the ingest.

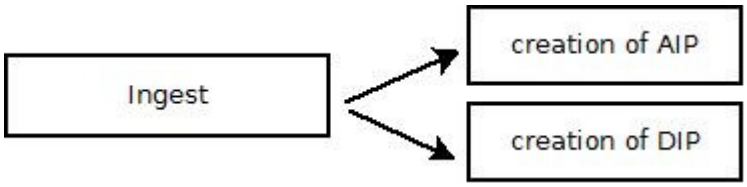

Figure 1: Creation of AIP and DIP during the Ingest

Users of the Federal Archives gain access to these collections anyway: They can ask for digital archival materials on demand or request specific information. At present users cannot browse our content online, as there is no connection between our online catalogue BASYS Invenio and the Digital Archives yet. This drawback only concerns born-digital content: The

Federal Archives already grant access to digitized archival content[10] via the online catalogue mentioned above.

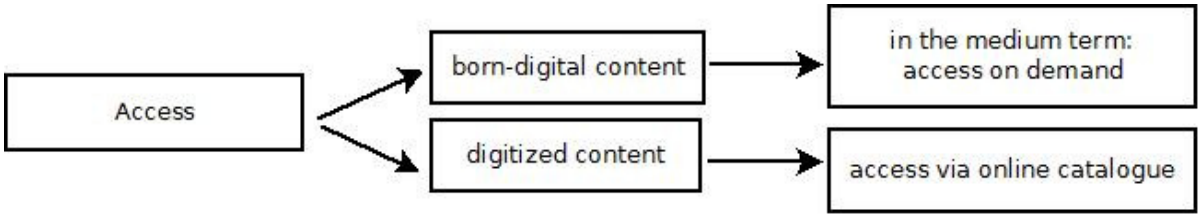

Figure 2: Access to born-digital content and digitized content

The Digital Archives ensures the data integrity, authenticity and validity of the digital archival content through technical and organizational measures, as will be exemplified in the following.

2.2 Data Model and system architecture of the Digital Archives

At present our six team members, five archivists and one ICT specialist, use the Digital Archive's applications exclusively. In the coming years a number of colleagues from several units within the Federal Archives will start using the system of the Digital Archives. Thus we made a good start and will strive further toward the creation of a user-friendly comprehensive system, the architecture of which is is as follows:

According to the OAIS information model (see above) the xml scheme "xBarch" forms the basis of our AIPs.[11] It provides for archival administrative metadata, content metadata, technical metadata, structural metadata and metadata for long time preservation; the technical part ("technische Daten") of the scheme is based on the Preservation Metadata Implementation Strategies (PREMIS), which were published 2003 and contain a list of recommendations concerning metadata for Digital Archives.

Archives have the responsibility of ensuring that the digital materials they have charge of remain understandable and usable as authentic copies. Hence we guarantee the link between the content and the information required for understanding and decipherment through the inclusion of the metadata within the AIPs as self-describing objects.

Obviously our first aim was to provide for a secure storage and the integrity of every bitstream ingested into our system as an important part of our digital preservation management (bitstream preservation). We have decided in favour of a hierarchical storage management (the product is called Grau ArchiveManager[12]). Accordingly we use a combination of large disk storage, which provide rapid access to recently saved AIPs, and magnetic tape storage for data which is accessed rarely. The software includes automatic error detection: If errors are detected the respective medium is marked as "unreliable" and further writing to it disabled. As a precaution four copies of each AIP are saved.

The next issue was the creation of a functional ingest component, which had been designed to be well adapted to record keeping systems. But the Electronic Document and Records Management Systems (EDRMS) or other computerised record keeping systems were not introduced in the government agencies as speedily as had been anticipated by the Federal Archives.[13] Instead there were other digital object types, which had to be integrated in the system, notably databases, digital media (image files, video files, audio files) and especially unstructured data (in shared folder file systems). Thus in 2013 we changed some components within the Digital Archives and implemented the PreIngest Toolset (PIT), which enables us to build SIPs out of unstructured data and databases.

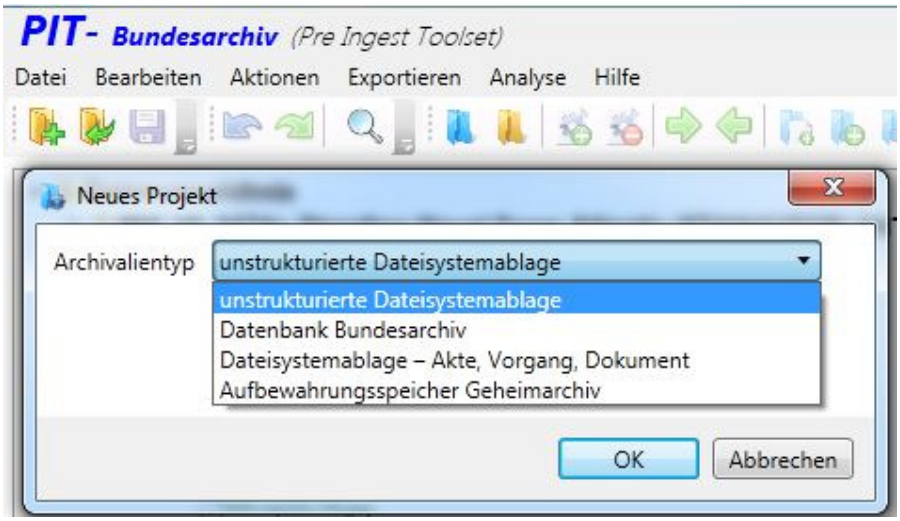

Figure 3: PreIngest Toolset (here: options for structuring)

Our long lasting preoccupation with the anticipated electronic records from EDRMS had some effect upon the tool's design as well as on our Xbarch scheme: We decided on the creation of three different options for structuring the unstructured data.

Provided that the data is patterned in a way which allows to form a trinominal structure (file/information package/document)[14], we are able to do this by means of the PIT. If the data is rather structureless we apply another structuring method ("unstructured data"), which permits more complex stratifications. There is a third option available, which enables us to form SIPs out of small databases. Thus we have integrated SIARD, which stands for Software-Independent Archiving of Relational Databases, into the PIT. SIARD was developed by the Swiss Federal Archives for the archiving of relational databases. It is based on international standards (amongst others Unicode and XML) and finding increasing use worldwide.

All these options have an impact on our Xbarch scheme. Depending on the type of record (unstructured data, file/dossier/document or database) chosen at the beginning of the structuring process, different metadata elements are filled in.

For the creation and management of the AIPs we use a software suite, which includes a consistent metadata management (DOXiS4 iECM Suite / SER). Hence we establish a tool for the digital preservation management.

2.3 Excursus: Digital Preservation Management in the Federal Archives

2.3.1 Introduction

Digital preservation poses challenges to archives worldwide, which can only be coped with by applying a comprehensive preservation management: Whereas the traditional preservation can resort to long standing reviewed conservation techniques, the methods of the digital preservation management have to be constantly adapted to rapid changes in the technical infrastructure.

Given this fact, the impact of certain aspects increases, especially the systematic monitoring of all technical processes within the Digital Archives and an information management, concerning all important issues of the Digital Archives.

Having covered one important aspect of digital preservation management from the start by the implementation of a special hierarchical storage management for bitstream preservation (see above), the Digital Archives began in 2011/2012, in conjunction with Hewlett-Packard (HP), to develop the browser-based tool „Digital Preservation Manager" (DPM), formerly

known as „Bestandserhaltungstool“, in order to ensure, that the digital archival materials remain permanently usable (logical preservation).

2.3.2 Digital Preservation Management within the Digital Archives

The Digital Archives establishes its digital preservation management in accordance with current national and international standards and best practices.[15] In future all implemented preservation plans shall arise from our preservation policy[16], the main aim of which is to preserve the integrity, authenticity, confidentiality and usability of the digital archival materials.

Thus the Digital Preservation Manager tackles three important tasks: Firstly it allows specifying the software and hardware environments as well as the organizational, financial and legal environments required to preserve the AIPs (information management). Secondly it enables us to identify the information packages in danger of becoming unusable due to technology changes or flaws within the organizational or legal environments, which pose a threat to the Digital Archives (risk assessment). Thirdly the tool makes it possible for us to plan the activities associated with preservation, for example implementing migration strategies, revising the preservation policy or the legal foundation and changing the organizational structure of the Digital Archives et cetera (preservation planning / preservation action). Thus the application enables our team to assess the risks for the Digital Archives (e.g. concerning a specific file format) and invent suitable preservation plans.

A digital preservation life cycle is formed:

Figure 4: The tool Digital Preservation Manager (here: slide “Bestandserhaltung” i.e. preservation)

The process starts with the “Status Quo”, where the framework requirements, objectives as well as technical, organizational, financial and legal structures of the Digital Archives are held. Sensitive matters or rather themes, which might become threatening, can be tagged. Afterwards one can set a deadline, which activates a reminder in the “Monitoring”. There the preservation planning is carried out: (Firstly) if the time-limits are exceeded, an alarm is triggered. (Secondly) if there is any evidence (integrated information sources, e.g. registries like the Pronom technical registry or selective disseminations of information like Google Alerts) that important issues threaten the Digital Archives, an alarm is triggered as well.[17] Based on all important information the preservation plans are designed (button “Planung”) and implemented in a controlled environment (button “Umsetzung”). Subsequently the whole process can be closed, all actions are documented (button “Historie”). An updated theme arrives at the beginning and the life cycle is completed.

Another most important functionality within the Digital Archives has been incorporated in the Digital Preservation Manager: As of late the tool supports normalization (i.e. format migration, button “Migration”) in order to reduce the number of formats the Digital Archives has to cope with, given that long-term access to each of these formats cannot be guaranteed. Thus our preservation policy implies the provision of a secure storage (see above) and the migration of records to new technical manifestations as two of our preservation strategies.[18] The latter imports the file format migration of text files into PDF/A (1a)[19], whilst preserving the original. Furthermore we plan to implement a file format migration of image files into TIFF. Audio and video files are currently preserved only in the original file format.

We don’t use the Digital Preservation Manager for the actual file format migration. The tool serves as a monitoring application as well as an order management, while the unpacking and migration of the AIPs is triggered within the Doxis Webcube (which is the web client of the product “DOXiS4 iECM Suite” from SER) and afterwards carried out within a tool named “rendition server” (HP/Adlib). The post processing includes error checking routines by means of the validation tools Jhove (characterization tool), Droid (format identification) and Pronom (technical registry). Recently we have been planning a stronger integration between the various functions of our system and the Digital Preservation Manager, so as to furnish for example a display of all file formats within the archival storage or to get information about detected errors by a connection with our storage management (the Grau ArchiveManager); by this means connecting the tool for logical preservation with the component for bitstream preservation.

For the implementation of different preservation strategies (e.g. file format migration, emulation et cetera) we can revert to several system environments, where all strategies are put to test before they are implemented. Therefore we will possibly enhance the Digital Preservation Manager further in order to use the tool as a comprehensive configuration management application.

3. Summary and Outlook

The history of the Digital Archives within the Federal Archives began with the acquisition of the data originating from the government agencies of the former GDR and the foundation of a specialised electronic records unit (which was not called “Digital Archives” back then) in 1991. At the beginning, our former colleagues, who were responsible for this data in that decade, had to muddle through, but they achieved a lot: They tracked down dearly needed information (code lists, metadata) without which it would have been impossible to read the records and did pioneering work on digital preservation. When the Digital Archives of the Federal Archives was established in 2006 one did not have to start afresh, but could resume their work. Nevertheless there was a lot of work to deal with: The goal was to establish a functional Digital Archives according to international standards (such as OAIS, PREMIS and METS).[20] Thus we have made great strides in the last years. By now we render most of the services described in the OAIS Reference Model and have implemented a pre-ingest procedure because of the huge amounts of unstructured data from the German government agencies. When the former unit responsible for electronic records ceased to exist in 2001 the Federal Archives had already been storing digital archival materials “with a special emphasis on data from agencies of the former DDR [GDR]”[21]. Today only roughly 10% of the content stored within the Digital Archives once belonged to agencies of the former GDR.

Nonetheless we have not fully accomplished the task yet: We will need some more time to implement fully fledged user services. Moreover we will work hard to develop the digital preservation management still further. Lately we have had to face the need for a special storage system for classified documents, which we are now establishing within the Digital Archives (and which has been affected the design of the PIT). This won’t be the last time we have to meet new requirements of the public body, which frankly is quite an engaging perspective.

There are many challenges archivists have to meet today, amongst others, increasingly larger volumes of electronic records in larger file sizes and in a variety of file formats as well as new forms of interaction and communication in the public sector, for example the increasing use of web application frameworks (e.g. Microsoft Sharepoint) and social media (e.g. Facebook, Twitter, Blogs) in the government agencies. Unlike the UK Government the Federal government has not yet recommend a “Cloud First” policy for the public sector to achieve better value for money in IT services and data storage.[22] Instead the Federal Government relies on a shared storage service, while trying to concentrate all Federal ICT services in Germany. As the German government agencies are now required by law[23] to establish Electronic Document and Records Management Systems (EDRMS) by 2020 the Federal Archives is concerned and has reacted: We have offered the wished for shared storage as a new service of the Digital Archives called Digital Intermediate Archives. Thus the Federal agencies won’t have to establish their own digital repositories and to resort to commercial cloud storages either. If this plan is carried out, the Government and its agencies will get a secure storage to preserve their digital information assets, while the Federal Archives will be able to make sure, that the digital heritage is saved for future generations.

Bibliography

Beagrie, Neil et al. (Ed.) (2014): How Cloud Storage can address the needs of public archives in the UK. Guidance on Cloud Storage and Digital Preservation, p. 5: www.nationalarchives.gov.uk/documents/archives/cloud-storage-guidance.pdf [accessed 7.8.2014]

Brown, Adrian (2013): Practical Digital Preservation. A how-to guide for organizations of any size, Lodon: Facet Publishing.

Hänger, Andrea/Lupprian, Karl-Ernst (2005): Archivierung elektronischer Unterlagen der Verwaltung. Aussonderungskriterien, Langzeitstabilität und Authentizität, in: Zeitschrift für Bibliothekswesen und Bibliographie 52/3-4, pp.137-142.

Hänger, Andrea (2006): Einleitung. Vom „hybriden Chaos“ zur elektronischen Vorgangsbearbeitung, in: Vorgangsbearbeitungssysteme in Bundesministerien (Materialien aus dem Bundesarchiv 18), Koblenz, pp. 7-15.

Hänger, Andrea (2008): Aussonderung und Archivierung elektronischer Akten - Vorstellung des Pilotprojektes des Bundesarchivs, in: Archive in Bayern 4, pp. 243-249.

Hänger, Andrea (2009): Der Stand der Einführung elektronischer Vorgangsbearbeitung in Bundesbehörden, in: Zwischen analog und digital - Schriftgutverwaltung als Herausforderung für Archive: Beiträge zum 13. Archivwissenschaftlichen Kolloquium der Archivschule Marburg, ed. Alexandra Lutz, Marburg, pp. 157-169.

Hänger, Andrea/Ernst, Katharina (2010): Ein System - zwei Lösungen: Digitale Archivierung im Bundesarchiv und im Stadtarchiv Stuttgart, in: Archive im digitalen Zeitalter: Überlieferung, Erschließung, Präsentation. 79. Deutscher Archivtag in Regensburg (Tagungsdokumentationen zum Deutschen Archivtag 14), ed. Heiner Schmitt, pp. 77-83

Leitfaden (2014) zur Erstellung einer institutionellen Policy zur digitalen Langzeitarchivierung, verfasst und herausgegeben von der nestor-Arbeitsgruppe Policy (nestor-materialien 18). http://www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/Materialien/materialien_node .html [accessed 18.06.2014].

Rathje, Ulf (1999): Digitale Datenbestände von Behörden und Einrichtungen der DDR im Bundesarchiv, in: Historical Social Research 24, 4, pp. 70-101. URN: http://www.ssoar.info/ssoar/handle/document/3166 [accessed 10.08.2014].

Rathje, Ulf (2003a): Archivierung von DDR-Daten im Bundesarchiv: Ein Rückblick auf zehn Jahre, in: Historical Social Research 28, 1/2, pp. 57-72. http://www.ssoar.info/ssoar/handle/document/3387, [accessed 11.08.2014].

Rathje, Ulf (2003b): Daten der Staatlichen Zentralverwaltung für Statistik der DDR im Bundesarchiv – Bewertung, Übernahme, technische Bearbeitung und Benutzung, in: Mitteilungen aus dem Bundesarchiv, 11/3, pp. 56-63. http://www.staatsarchiv.sg.ch/home/auds/06/_jcr_content/Par/downloadlist_2/DownloadListPar/download_9.ocFile/Text%20Rathje.pdf, [accessed 11.08.2014].

Schroeder, Kathrin/Huth, Karsten (2007): Projekt „Aufbau eines Digitalen Archivs“ im Bundesarchiv: Werkstattbericht über den „individuellen“ Umgang mit Massendaten, in: Erfahrungen mit der Übernahme digitaler Daten: Bewertung, Übernahme, Aufbereitung, Speicherung, Datenmanagement (Tagung des Arbeitskreises Archivierung von Unterlagen aus Digitalen Systemen 11), ed. Katharina Ernst, Stuttgart, pp. 42-46.

Schroeder, Kathrin/Huth, Karsten (2009): Das Metadatenkonzept des Digitalen Archivs des Bundesarchivs, in: Der Archivar 62/3, pp. 248-254.

Schürer, Kevin (Ed.) (2001): Better access to electronic information for the citizen. The relationship between public administration and archives services concerning electronic documents and records management (European Archives News, Supplement V), Luxemburg.

The National Archives (2011). Digital Preservation Policies. Guidance for archives, http://www.nationalarchives.gov.uk/archives-sector/developing-a-strategy.htm [accessed 18.06.2014].

Wettengel, Michael (1993): Abteilung II: System zur Archivierung maschinenlesbarer Daten im Bundesarchiv, in: Mitteilungen aus dem Bundesarchiv 1/2, pp. 70-72.
Wettengel, Michael (1995a): Archivierung maschinenlesbarer Datenbestände im Bundesarchiv, in: Historical Social Research 20/4, pp. 123-127.
Wettengel, Michael (1995b): Maschinenlesbare Datenträger: Zusammenstellung archivrelevanter Normen und Standards elektronischer Speichermedien, in: Der Archivar 48, pp. 461-471.
Wettengel, Michael (1997): Technische Infrastruktur für die Archivierung von digitalen Datenbeständen, in: Mitteilungen aus dem Bundesarchiv 5/1, pp. 8-15.
Zahnhausen, Vera (2012): Das Digitale Archiv des Bundesarchivs: Ein aktueller Überblick, in: Mitteilungen aus dem Bundesarchiv 20,1, pp. 31-35.
Zahnhausen, Vera (2013): Überlieferungsbildung zwischen analog und digital: Erfahrungen bei der Übernahme von digitalem Archivgut, in: Mitteilungen aus dem Bundesarchiv 21,1, pp. 44-50.

Notes

[1] The State Archives of the former German Democratic Republic were merged with the Federal Archives in 1990.
[2] Cf. (Rathje, 2003a, pp. 57-72). S. (Schürer, 2001, p. 43). Cf. (Rathje, 1999, pp. 70-101). Cf. (Wettengel, 1995a, pp. 123-127).
[3] Cf. (Rathje, 2003b, pp. 56-63).
[4] For comparison only: The National Archives and Records Administration (NARA) Electronic Records Archives project was started in 1998, whereas the UK National Archives (TNA) began to develop its digital repository in 2001. (Brown, 2013, p. 66).
[5] Concerning the history of digital preservation compare the very concise account in: (Brown, 2013, pp. 9-12).
[6] Cf. (Hänger/Lupprian, 2005, pp.137-142). (Schroeder/Huth, 2007, pp. 42-46). (Hänger, 2008, pp. 243-249). (Hänger, 2010, pp. 77-83). (Zahnhausen, 2012, pp. 31-35). (Zahnhausen, 2013, pp. 44-50).
[7] Unlike the TNA: http://www.nationalarchives.gov.uk/information-management/, [1.8.2014].
[8] Like the National Archives of Australia. [1.8.2014]: http://www.naa.gov.au/records-management/agency/digital/EDRMS/index.aspx.
[9] While building the submission information packages (SIPs).
[10] Which is stored within a so called „Digitales Magazin“, which simply means digital repository.
[11] Cf. (Schroeder/Huth, 2009, pp. 248-254).
[12] Cf. www.graudata.com/english/ArchiveManager/function [6.8.2014].
[13] Cf. (Hänger, 2006, pp. 7-15). Cf. (Hänger, 2009, pp. 157-169).
[14] Corresponding to the principles of archival description at the Federal Archives.
[15] For example OAIS, “NESTOR-Kriterien für vertrauenswürdige digitale Langzeitarchive”, „Digital Preservation Planning“-concepts [2012-2014] of the National Archives of the United States, New Zealand, Australia, Sweden, the United Kingdom as well as the National Libraries of the Neatherlands and Great Britain. See (Brown, 2013).
[16] This will be published within the following months.
[17] This function is not implemented yet.
[18] Cf. (Leitfaden, 2014, p. 3). See (The National Archives, 2011, p. 8): „The digital preservation policy must identify a digital preservation strategy and its owner(s). This reinforces the strategy and the preservation activities it defines, whilst allowing the policy to remain independent and not have to be updated when change is made.“
[19] Like the National Archives of Australia (tools Xena) and Archivematica.
[20] Untill 2001 “the general hardware platform” was “a Solaris (SUN) system, with Unix as the operating system. This connect[ed] to a number of tape drives and a PC.” There was “also a workstation with a specialist tape drive facilities. The Thomson software package [was] used for copying and analysis.” (Schürer, 2001, pp. 18). The preservation media used back then were open reel tapes (all types), DAT and Exabyte cassettes and CD-ROMs. Preservation formats were ASCII and EBCDIC, for “office documents” TIFF 6.0, and CCITT/TSS. Cf. (Wettengel,1997, pp. 8-15). Cf. (Wettengel, 1995b, pp. 461-471). (Wettengel, 1993, pp. 70-72).
[21] (Schürer, 2001, p. 17).

[22] Cf. (Beagrie, 2014, p. 5).
[23] The Law on electronic government passed the parliament (“Bundestag”) in 2013.