Exploring Friedrich Kittler's Digital Legacy on Different Levels: Tools to Equip the Future Archivist

Jürgen Enge
University of Art and Design (FHNW) Basel
Freilager-Platz 1
4023 Basel, Switzerland
+41 61 228 41 03
juergen.enge@fhnw.ch

Heinz Werner Kramski
Deutsches Literaturarchiv Marbach
Schillerhöhe 8–10
71672 Marbach, Germany
+49 7144 848 140
kramski@dla-marbach.de

ABSTRACT
Based on the example of Friedrich Kittler's digital papers at the Deutsches Literaturarchiv Marbach (DLA), this paper explores digital estates and their challenges on different logical levels within the pre-archival analysis, documentation and indexing process. As opposed to long-term digital preservation procedures, which are set about afterwards when relevant digital objects have already been identified, this process starts shortly after physical material (computers, hard drives, disks…) is delivered to the archive and has been ingested and safeguarded into volume image files. In this situation, it is important to get an overview of the "current state": Which data was delivered (amount, formats, duplicates, versions)? What is the legal status of the stored data? Which digital objects are relevant and should be accessible for which types of users/researchers etc.? What kind of contextual knowledge needs to be preserved for the future? In order to address these questions and to assign meaning to both technological and documentation needs, the digital analysis tool "Indexer"[1] was developed [3]. It combines automated, information retrieval routines with human interaction features, thereby completing the necessary toolset for processing unstructured digital estates. It turns out however, that intellectual work and deep knowledge of the collection context still play an important role and must work hand in hand with the new automation efforts.

Keywords
Digital estates; digital papers; personal digital archives; digital analysis; file format identification; pre-archival process; semi-automated indexing; appraisal.

1. INTRODUCTION
The collections of the German Literature Archive (Deutsches Literaturarchiv – DLA) bring together and preserve valuable sources of literary and intellectual history from 1750 to the present day. Around 1,400 conventional papers and collections of authors and scholars, archives of literary publishers and about one million library volumes still make up the bulk of the collections. With the emergence of text processing and computer-assisted work for writers, authors and publishers, digital documents surely belong more and more to the field of collection of literary life and German-language contemporary literature.

With regard to digital unica – that usually remain unpublished and restricted to a single data carrier – a memory institution bears extraordinary responsibility for their long-term preservation, since per se no cooperative or redundant collection and indexing can be undertaken.

When DLA first began processing the digital estate of Thomas Strittmatter (1961–1995), it was one of the first memory institutions in the German-speaking world that needed to develop a workflow for digital unica [14, 77; 13]. Since then, 281 data carriers (almost exclusively 3.5"- and 5.25" floppy disks) from approximately 35 collections were saved, and roughly 26,700 files converted into stable formats.

With the exception of Strittmatter's Atari and F.C. Delius' Macintosh, only data carriers were acquired during this phase, but no complete computer environments. Often, disks were discovered incidentally while examining the conventional material, rather than deliberately acquired. Our priority was to conserve the texts as objects of information independent of their respective carriers. The two PCs were displayed in the Museum of Modern Literature (Literaturmuseum der Moderne – LiMo), but only as museum exhibition pieces, not as functional working environments in the sense of [5].

The digital estate of Friedrich Kittler (1943–2011), which was acquired in spring 2012 without any technical pre-custodial preparations, goes beyond the scope of previous procedures, both quantitatively and qualitatively. Thus, it became necessary to explore new options: Digital analysis tools and automated work routines have been brought into focus, in order to make the yet-unknown amounts of data manageable.

Friedrich Kittler was one of the most famous and important German media theorists and literary scholar. His impact on humanities in general and media studies in particular is of growing interest due to technological and methodological reasons. Since Kittler's media archeological merits have derived to a great extent from his practical experiences in programming, it seems comprehensible that his intellectual legacy can only be understood and/or reconstructed by accessing both his theoretical work (books, articles, documented presentations) and his digital programming experiments. Whereas parts of the first were mostly published during his lifetime, the latter is basically hidden on nine hard drives, 104 optical disks, 648 floppies, etc. – hereinafter referred to as "Kittler's digital estate". Both are supposed to be (re-)edited in the now compiled Kittler edition.

Kittler bequeathed collected source codes as well as modifications of his own software and hardware configurations. Among the rather "idiosyncratically" [4, 14] structured data are thus "handwritten" codes, like Kittlers 15 years spanning computer-based study of Markov chains, which "one might say, [forestall ...] Digital Humanities, since they constitute computer-based text analysis" [4, 12].

The wish to encounter scholarly pieces in their original, immediate environment and folder structure of Kittler's personal computing working place made it necessary to show

[1] In jest we call the Indexer "Ironmaiden": "Intelligent Read-Only Media Identification Engine" or "Intelligent Recursive Online Metadata and Indexing Engine" but the official name simply is "Indexer".

utmost restraint in excluding data from future access. Even more so since Kittler routinely worked with root privileges, thus having and using writing authorizations everywhere. In other words, it was very important to find ways to make Kittler's source codes accessible – especially within their immediate local context at the original file location and position in the system directories. Therefore, all files from hard drives, most of the readable disks and about all optical media were examined. Only obvious "mass products" (such as CD-ROMs attached to the German computer magazine "c't") were only registered, but not copied.

Whereas from a technical point of view the heterogeneity of different file formats and the sheer mass of 1.7 million files were demanding, regarding semantic challenges it soon became clear that human interaction and decision making was indispensable. At the same time even these intellectual decisions had to be formulated in a rather formal way so that they could be applied to whole groups of (technically) similar data. Hundreds or thousands of files were just too much to analyze manually and the risk of publishing semantically restricted files was just too big.

2. IDENTIFYING AND DOCUMENTING

Since technological and semantic challenges of Kittler's digital estate did increase the documentation needs, implicit information had to be made explicit. Hidden knowledge had to be documented and assigned to its host components for enabling future investigations. As opposed to approaches which focus primarily on the content part of the digital estates and/or the raw files, the pre-archival indexing and appraisal processes meant in our case adding and keeping contextual information, too. Contextual information might be attached to the physical/hardware carrier (traces of handling) or conventions in naming or storing information at dedicated places, so careful documentation is recommended. Keeping track of this information supports later access regulations.

In his presentation, Christopher Lee 2012 defines eight "levels of representation" of digital resources [7, 7]. In contrast, we propose introducing an additional "level -1":

– Hardware (primarily as a museum object).

The sequence of our six levels is roughly related to the order of treatment. Combining a rather documentarian approach with institutional and operational needs in the pre-archival indexing process, we suggest furthermore at least five chunks of information entities:

– hard disks and data carriers (in terms of physical computing or storage media)
– (raw) disk images, which provide an important archival backup copy
– filesystems, indicating information about the used operating systems
– raw files, which contain the content/data
– context(ual) information, which is subsequently generated (in terms of virtual layer).

The following considerations start with the rather documentarian part which focuses on the first three levels: hardware, hard disk and data carrier, and image backup.

3. HARDWARE

The relevance of the hardware level again becomes apparent when considering the case of Kittler's estate: During April 2012, the DLA first received two older tower PCs from Kittler's estate, both of which had not been used for some time (his current PC was initially kept in Berlin, as a hardware reference for Kittler programs, and was at later date forwarded along with additional old laptops).

At first, from the perspective of conventional preservation raises the issue of cleaning the soiled and dusty hardware components. Due to the danger of carrying mold spores into the magazines, it was decided to remove loose dust, but to keep attached traces of grime and liquids as authentic signs of usage. For a reset button strewn with pen and pencil marks is a testimony of how often its adventurous user had to irregularly reboot his computer. Even after a complete migration and emulation of all digital objects, the hardware retains the nimbus of an original and potential exhibit.

During this early phase, it has proven valuable to decide on distinct (though not always chronologically correct) labels for the computers ("PC1", "PC2") and to keep a dossier with many photographs from the very beginning, in order to document and keep track of the growing amount of hardware information.

PC1 was brought to the archive without a hard drive, was non-functional and so was documented via visual examination only. With the help of live boot media (for example Xubuntu 8.10, which had to be used due to the limited RAM equipment) and tools like "lshw", "lspci", "lsusb", "hwinfo", "inxi" etc., the hardware configuration of PC2 and later, functional computers was analyzed.

Figure 1. One of Kittler's old PCs (Pentium III, ca. 2000) showing heavy signs of usage on the reset button.

The inspection and analysis of the hardware required substantial employment of personnel, as well as profound IT knowhow, preferably with Linux distributions and hardware components of the period of use (such as SCSI hard drives and controllers). On the other hand, standardized live media and hardware diagnosis tools are available, which allow for a precise and fast overview. Apart from purely technical work, information about the usage context has to be collected, as this may influence the prioritization of tasks. For example, it became necessary to contact Kittler's former colleagues to learn his login password.

4. HARD DISKS AND DATA CARRIER

Very often data carriers are physically contextualized by the technological context in which they occur: a build-in hard drive fulfills different functions in most of the cases than a portable one. One might also differentiate semantically between a rather active usage of data carriers, which are continuously in use and thus integral part of the working process, and passive usage, in which data carriers are accessed only temporarily. Passive data carriers instead are often used for transporting data through time and place; they contain data which the owner kept with him/her for presentation or backing-up reasons, which might indicate a certain kind of relevance.

Since Kittler was a heavy smoker and a lot of dust settled down on data carriers stored under non-optimal conditions over the

years, all volumes first entered the conservation and restoration team of DLA, which subjected the storage media to professional cleaning.

Before any further processing could be made, it had to be ensured that the write-protection of floppy disks was active. Because of the wide range of filesystems used on disks (including many "ext2" formatted ones), all reading operations have been carried out on Linux.

In a first reading step, all floppy disks were processed by a long command line which recorded – besides other technical metadata – the filesystem type and the change time of the most recent file contained on disk. This date was then temporarily attached to each volume by sticky notes and allowed manual re-consolidation of scattered disks to a joined set, for example a particular backup. This formation of groups could usually be confirmed by the disks' look and feel (make, labeling, signs of usage).

The cleaning and sorting was followed by a carefully designed labeling process, where internal identifiers were assigned to all hard disks and removable media.

- The acquired hard drives were distinctly labeled "hd01", "hd02" etc., which is to be understood as a numerus currens without chronological significance. A hierarchical attribution of internal hard disks to computers was not possible, since they were often either installed and functional, installed but not connected or completely separated with no way of determining which PC they belonged to.
- The naming of the contained partitions was largely based on another pattern, independent of the naming conventions of the running operating system. Other names for data carriers were defined as follows:
- fd001 etc.: floppy disks, disks
- od001 etc.: optical disks, CD-ROM, CD-R, CD-RW, DVD etc.
- xd001 etc.: external files: File collections on other external data carriers, e.g. on USB hard drives of the DLA.

The labels were written with a permanent marker on labeling boxes on cable ties, or on (mostly the backside) of the carriers themselves. For the labeling of black floppy disks, using "Posca" markers with water-solvable white pigment ink, the kind of which is also used by conservators, has proven successful.

These labels also served to create file names for sector images and listings and simplified the administration in internal lists that could later be imported into the Indexer. However, these labels are not identical to the archive's accession numbers, since those had not yet been assigned at that point.

Similar to hardware, inspecting, analyzing and possibly consolidating the data carriers required both substantial employment of staff and profound IT knowhow. However, via scripts and standard Linux tools ("mount", "ls" etc.) the analytical steps for disks can be conveniently automated. In Kittler's case, who archived numerous self-made copies of MS-DOS programs and operating systems on disks, knowledge of 1990s software is helpful for identifying and classifying these disks. Susanne Holl has shown that the frequency and occurrence of specific files on active and passive data carriers can reveal interesting information regarding relevance: "it is an interesting piece of information," she states, "that machine.txt was saved 22 times, itinerating through all hardware upgrades, from hard drive to floppy to hard drive to optical disk to hard drive" [4, 8].

Furthermore, close cooperation with one of his colleagues has been invaluable because she could identify many data carriers as Kittler's "writings" in the narrow sense of the word, which influenced the chronological order of further steps.

5. IMAGE BACKUP

Although DLA only began in 2014 (with the acquisition of Kittler's most recent PC) to actively use tools from the BitCurator distribution, almost from the beginning in 2003 it followed a strategy highly recommended by the BitCurator project: to conservate media volumes as a one-to-one copy into sector images, the "cornerstone of many forensics methods" [8, 27]. Recovery and analysis of deleted files is not part of DLA's standard workflow, but based on these images, it would at least be possible in cases of special need.

Figure 2. Running BitCurator (live medium) on Kittler's last PC (Intel Core i7-2600K, 2011).

In general, sector images are most qualified to preserve technical metadata of filesystems (update time stamps, user information etc.). Moreover, they can be directly integrated as logical drives (read-only) in virtual machines or emulators (see Figure 7).

Sector-wise copying of floppy disks could not be carried out with the previously used, custom-made windows tool "FloppImg" [13], because of the large amount of ext2 and other filesystems not mountable on Windows. A Linux script was used instead which calls the tool "ddrescue" and hence works well with deficient media.

244 disks out of a total of 648 were initially not considered during this work step, because they were obviously industrially produced installation disks for operating systems, drivers or application programs (MS-DOS, Windows 3.x, SCO Unix, Gentoo Linux) or 1:1 copies of the same. Their backup into the DLA software archive, which is established independently of Kittler and could be relevant to future emulations, is still pending. Whether these data carriers can be counted among Kittler's digital estate in the narrow sense, is open to debate. (When installed on his hard drives and theoretically executable, they certainly do, as they form his working environment.) But when in doubt, disks labeled either by handwriting or by typewriter were considered relevant and thus copied. Some disks were simple empty and not in use. However, disks that were apparently empty, but had handwritten labels were examined more closely using "Kryoflux". Out of 404 interesting candidates, it was in 119 cases not possible to create mount- and usable sector images. Therefore, the failure rate of Kittler's disks (the oldest ones date from 1987) amounts to 29.5%

CD-Rs instead were converted into .iso-files by the c't tool "h2cdimage" which creates partially usable images from

deficient volumes like ddrescue [2]. In contrast to common copy programs it will not continue reading in deficient sectors without any further progress, so that the drive will not degrade from continued reading attempts.

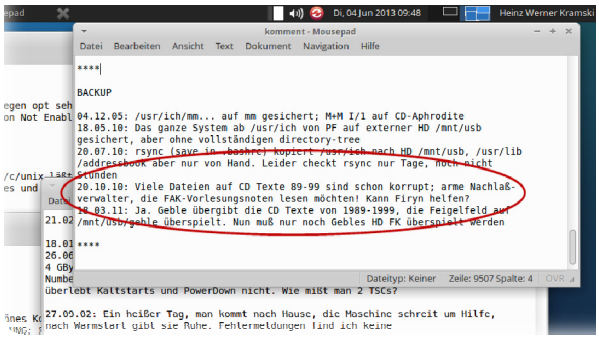

Figure 3. “Arme Nachlaßverwalter” (Poor curators)...

Regarding a central CD-R, Kittler says in the file “komment” (a kind of technical diary) “20.10.10: Many files on CD Texts 89–99 are already corrupt; poor curators who want to read FAK lecture notes!” [6]. It is remarkable to be addressed from the past in such a way. It is also remarkable how of all things, it was Kittler’s beloved c’t that helped save an unexpectedly high portion of backup CD-Rs that he had already dismissed as “broken“ during an interview [10].

Out of 104 optical data carriers, 82 were temporarily ruled out as mass-produced ware and installation media. Out of the remaining 22 self-burned CD-Rs, only three could not be flawlessly copied. However, it was possible to later mount them.

Hard drive partitions were at first also created using Linux scripts and ddrescue. From 2014 onwards, “Guymager” in “dd” mode (without file splitting) was used. Regrettably, there was an unreadable partition on a 2 GB SCSI drive.

Besides the principal difficulties of selecting relevant files for file format migration and for further editing, real technical problems arose in the attempt to store original files from hard disk partitions and optical volumes on the standard file servers of the DLA (as it was previously possible with the floppy disk inventory):

1. A digital estate is stored on the file server with an extensive path named after its holder with systematically labeled subfolders according to the processing state (see [12]). If original files to be stored have their own, deeply nested path hierarchy, the allowed path length of the operating systems involved might be exceeded.
2. Today’s virus scanners often impede the copying of original files contaminated with old (MS-DOS) viruses.
3. DLA’s standard file server does not support the original case-sensitive file names (e.g. Makefile vs. makefile) when serving Windows-based clients.
4. Reading errors often prevent file-by-file copying of original media.

It is possible to overcome all these limitations by mounting disk images, but then an appropriate presentation tool is needed. The Indexer therefore not only is required for full text indexing and MIME type analysis (see section 7), but also serves as a document server which preserves the authentic path information. However, the main motivation for developing and applying the Indexer remains the fact that 1,7 million files cannot be assessed by our colleagues in the archive without prior technical preparation, while at the same time, all technical measures must concentrate on a selection that can only be made through intellectual assessment. An implicit decision of relevancy, as it was possible in case of floppy disks, is bound to fail, when it comes to the enormous amounts of data contained by hard disks.

Although the primary reason for image copies are archival needs (backup, protection of the original source), they also offer a starting point for the indexing process, which can only start when an accessible filesystem is available.

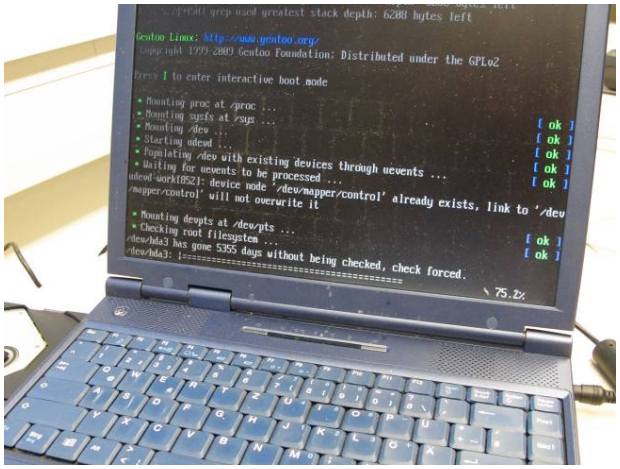

Figure 4. Why you should do disk imaging before anything else: “/dev/hda3 has gone 5355 days without being checked, check forced”.

For disk imaging there is very good tool support and long running copy or checksum jobs can easily be done on the side. Still, all steps have to be carefully monitored and documented, so IT knowhow is of advantage. However, as soon as data carrier identification has taken place and more detailed task schedules can be prepared, producing specific disk images can be delegated.

6. ANALYZING

Regarding the previously mentioned chunks of information, the analyzing part of the information retrieval starts at the filesystem level. It is followed by the raw files themselves and ends with observations regarding the contextual information. The result of the analysis of the filesystem are intentional statements because (at least parts of) the filesystem contain information about working process and conventions of the author: “We constantly seek not an artificially imposed classification by subject, but authentic pattern recognition of media in their archival order” [11, 112]. Kittler, for example, used several operating systems in parallel, including MS-DOS, SCO-Unix, early Windows versions and later primarily Gentoo Linux, which identify themselves due to their file structure. Furthermore and as already stated, he preferred working as “root” on Linux, bypassing administrative limitations normally applied to standard users. His standard working directory was not the commonly used subfolder of “/home”, but “/usr/ich” instead. At first glance, Kittler seems to place himself on one level with system directories under “/usr” in the filesystem hierarchy. It is more likely, however, that he simply continued a convention of his earlier SCO Unix, which did indeed place user directories under “/usr”. Still, the naming of his user account as “ich” (Me) certainly shows that he did not consider himself “one of several” users of his computers.

Inside his working directory a semantic order is largely missing, since he organized his files based on their character set: ASCII (“,asc”), Latin9 (“.lat”), UTF8 (“.utf”) [1, Min.: 13.50f]. Also, the usage of non-standard file extensions made an automated MIME type identification useful.

7. FILESYSTEM

Independent of Kittler's case, information of the filesystem comes in general close to classical cataloging information as far as author, title, date of creation, format etc. are recorded.

As preparation, the created sector images were made available to the Indexer VM via a read-only NFS share. There, they were mounted as loopback devices ("basepath" in table 1). To be able to use hundreds of these devices, a kernel parameter had to be raised. There was a highly specialized IRIX filesystem (XFS using "version 1" directories), for which current Linux systems no longer provide drivers. However, this could be mounted using a very old version of Ubuntu (4.10 with kernel 2.6.8) and copied on ext3, which, in this special case, provided the base for further steps.

From the documentation described in section 2.3, a list was loaded into the Indexer which assigned a unique session ID and a short description to every image (see Table 1).

For collecting and producing technical metadata, the Indexer first reads the ID of the archiving sequence (sessionid) specified on the command line for a particular image container and Indexer run. Then for each (recursively detected) filesystem object a distinct file identification number is generated (fileid), which refers to this specific indexing session. Another ID (parentid) identifies the folder, in which the directory entry is filed, and finally the file or folder name referred (name). The path of the directory entry is documented (path) as well as the basic type (filetype), for instances such as "file", "directory", "reference", the size of the file (filesize), and a checksum (sha256), which can be used for authenticity verification purposes.

Table 1. Session table of the Indexer (simplified excerpt).

sessionid	name	basepath	localpath	...
2001	hd01-p01	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/hd/hd01/p01	/u01/fk/hd/	
2002	hd01-p02	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/hd/hd01/p02	/u01/fk/hd/	
3001	od001	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/od/od001	/u01/fk/od/	
3002	od002	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/od/od002	/u01/fk/od/	
4001	fd001	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/fd/fd001	/u01/fk/fd/	
4002	fd002	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/fd/fd002	/u01/fk/fd/	

...	group	bestand	description	solrpath	
	hd	kittler	Partition 0,4 GB vfat, ca. 20040000, 1. Partition auf hd01 (IBM Deskstar, 32 GB, IDE) aus PC2	/solr/kittler	
	hd	kittler	Partition 15,7 GB ext3, ca. 20030000, 2. Partition auf hd01 (IBM Deskstar, 32 GB, IDE) aus PC2	/solr/kittler	
	od	kittler	CD-R iso9660, ca. 20010820	/solr/kittler	
	od	kittler	CD-R iso9660	/solr/kittler	
	fd	kittler	3,5" vfat, ca. 19900300	/solr/kittler	
	fd	kittler	3,5" vfat, ca. 19900300	/solr/kittler	

Later this is also double checked with entries of the National Software Reference Library (NSRL) of the American National Institute of Standards and Technology (NIST) in order to identify registered files of common software packages [9].

Furthermore, the date/time stamps when files were changed (filectime) or last accessed (fileatime) are of great importance. Care must be taken here to prevent unintentional modifications to the time attributes, so all containers strictly may not be mounted in write mode. Last but not least, all information of the Unix-call stat() ("stat") and the indexing time and date ("archivetime") are documented.

For storing this basic information and in preparation of the later full-text index, the Indexer maintains a directory of all filesystem objects and their technical metadata in a MySQL database. Metadata created during the information retrieval, as well as the information on the access path is stored beside the record. (The importance of the original path is emphasized by the implemented quotation routine, which displays an APA-like reference for citation). The naming convention of the session ID allows the administration of different filesystems/different estates or groups of objects. To uniquely refer to a single file a combination of sessionid and fileid is recommended.

During the first run, a copy of each file also is written into a balanced cache folder ("localpath" in table 1), so the image containers do not need to be present all the time. This also overcomes most of the limitations of common file servers outlined in section 5 and allows providing file links to the user without access to the archived sector images.

8. RAW FILES

Since the 'raw files' are supposed to contain the content of information itself, their analysis is of special importance. The iterative identification cascade of the Indexer analyzes the data step-by-step and optimizes the identification quality. Since every file identification tool has its own particular qualities and shortcomings, the Indexer combines different software tools. The list can also be changed, replaced or upgraded at any time. The varying results derive from different recognition algorithms and -databases within the single tools. Since contradictory statements can occur, the Indexer treats all results as equal, so that the user has to decide which information he or she would trust.

Among the mandatory tools the following software packages are of special importance: "Libmagic", which creates the initial list of files and tries to identify MIME type and encoding, and "gvfs-info", which has similar capabilities, but can sometimes deliver different results.

Highly recommended is furthermore "Apache Tika", which extracts not only the MIME type and encodings, but also the full text in case of texts. Extracted full texts are compressed with "gzip" to save cache space. "avconv/ffmpeg" is then used for extracting technical metadata from files, which "gvfs-info" has already identified as time based media (MIME type "video/*" or "audio/*"). "ImageMagick" is finally consulted for analyzing image- and PDF-data, of which it creates thumbnails. These thumbnails are used as preview images in the user interface.

In addition, "Detex" is useful for extracting the content (text) from TeX-files (MIME type "text/x-tex") by removing the TeX-commands. "Antiword" extracts full text from older Word-files (MIME type "text/application-msword), and "xscc.awk" extracts comments from the source code. The NSRL (locally imported into a Berkeley DB for performance reasons), which was already mentioned, is used for identifying software, which was not modified but only used by Kittler. The "md5sum" creates a checksum in one of the required formats, when matching against the NSRL is done.

The Indexer's core is a "SOLR" full text index. It collects the results of the iterative identification cascade in a separate, searchable index. This is mainly for performance reasons, but it also provides an autonomous subsystem, which is independent of the indexing and MySQL infrastructure. The full-text index

itself is made accessible through a web-based user-interface, which enables search and information retrieval.

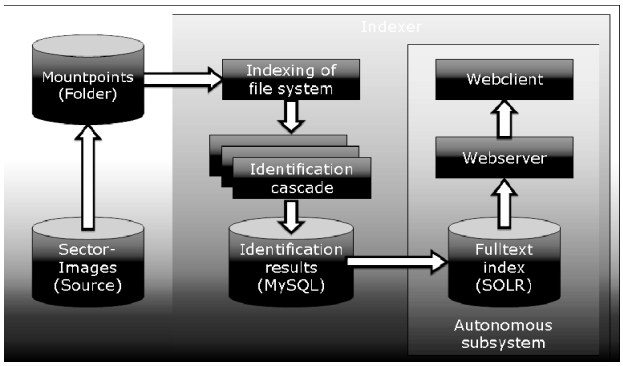

Figure 5. Indexer system architecture.

The simplified scheme above shows the overall system architecture of the indexer: Due to the large data volume, the Indexer runs were time-consuming and had to be gradually initiated and monitored. However, this effort is very much worthwhile: The knowledge gained through the automated MIME type analysis can hardly be overstated, since the estate is, from a traditional perspective, still unindexed. For example, a manual inspection might have classified word files with the extensions .doc, .DOC, .txt, .TXT, .dot, .DOT etc. as relevant for further investigation and possible migration of file formats. Unconventionally-labeled word files such as "*.vor" (presumably "Vorwort", preface) or "*.liz" might have escaped notice altogether.

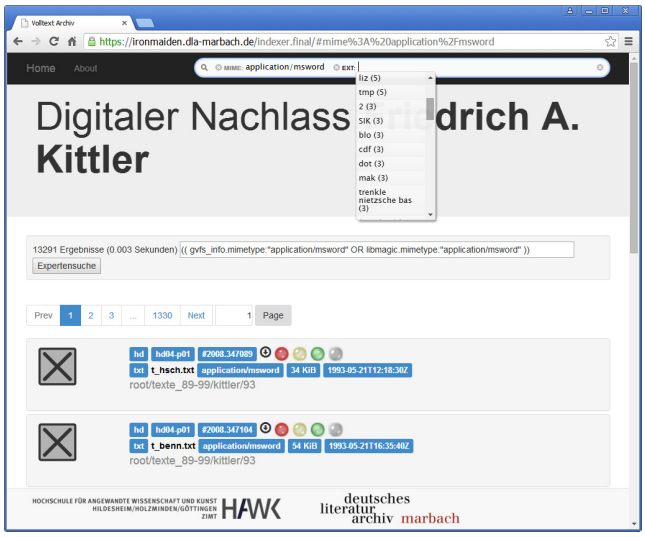

Figure 6. Searching for unusual MS-Word file extensions.

It must be noted however, that DLA has currently just completed the bitstream preservation work and did not yet enter the stage of systematic file format migration. Besides MS-Word, Kittler mainly used Emacs for text editing, so in the areas of scientific papers and source code, his digital estate should not impose too much future problems.

One notable exception are KWord files (".kw") for which no known migration tools seem to exist – even the direct successor, KDE's "Calligra" suite is unable to import the older, proprietary (pre-2005) ".kw" format. In a singular, important case, a Kittler Linux machine was brought to life again as a virtual machine and allowed to save these documents as ".rtf" files for further processing. But in general, virtualization (or emulation) currently requires too many manual arrangements to be part of an efficient standard workflow and will be addressed in particular by the planned edition of Kittler's collected writings, in whose edition plan a part for his own software projects is explicitly included.

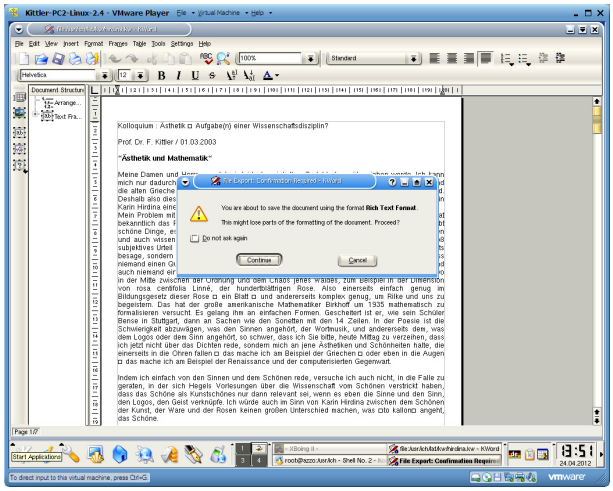

Figure 7. A Kittler VM running KWord version 1.2.1.

9. CONTEXTUAL INFORMATION

Beyond the technical analysis of data (indexing cascade), additional options for filtering are required. Since personal computers tent to contain private content (some may even be locked for 60 years by arrangement), information which touches third party personal rights (e.g. evaluation files), collected materials etc. withdrawing access rights from documents is essential. In case of DLA, suspending data is subject only to specific security measures and can only be imposed or removed by the administrators or the heiress. Questionable content can be added with a disclaimer, which informs the user that the data can't be accessed due to further specified reasons. The instance or file can thus still be referred to in the pre-view. Via a self-explaining 'traffic light' system access-rights can be visualized and changed.

Table 2. Indexer access levels

- Indexer record is unlocked (visible)
  - Show technical object ID only
  - Show metadata only
  - Show metadata and content, show fulltext search results in multi-line context, allow download (on campus)
  - Undefined, needs review
- Indexer record is locked (invisible; visible only to administrators)

Whereas withdrawal of usage rights can only be triggered by defined users who obtain specific editor rights and/or authorized scholars, locking off specific files, all other rights can only be set by administrators.

To execute mass classification which follows this scheme, formal rules have been created, which use server-side scripts. Among the applied routines are the following logical operations:

– Blur all thumbnails and set access level "Red" for all files having "mimetype:'image/jpeg'". This causes that all private photos get protected; however a great number of unproblematic images gets hidden as well. Another example of this type may be: Set access level "Yellow" for all files having "application/mbox" or "message/rfc822". This

protects the content of all incoming or outgoing emails. These rules can easily be applied by some SQL statements as the MIME type is already known, so the degree of automatization is high.

– Set a specific access level for selected folders or file names (which are known) to be especially problematic ("Red" or "Yellow") or especially unproblematic ("Green"). This only works based on lists created manually by Kittler's widow, to whom the inventory is well known. It also works only because Kittler's use of folder and file names remained quite stable over the years and through different (backup) volumes. However, manual work involved in this step is high and the risk of missing some problematic files or folders cannot be eliminated.
– Set access level "Green" for all files found in the NSRL. This is easy to do, but unfortunately only covers the less interesting files. (At least, it reduces the amount of files to be processed further by roughly a third or 570,000.)

Setting and checking access levels is still ongoing work.

Another type of contextual Information, which follows the principle of metadata enrichment, is implemented for future use with a checkbox system. Scholars and/or editors can classify entries according to DLA's standard classification with respect to the content. Additional features like a discussion forum might be appropriate to add in the midterm.

Currently filtering options are primarily meant to support the preliminary classification process or to filter data which is not yet meant to enter the public sphere.

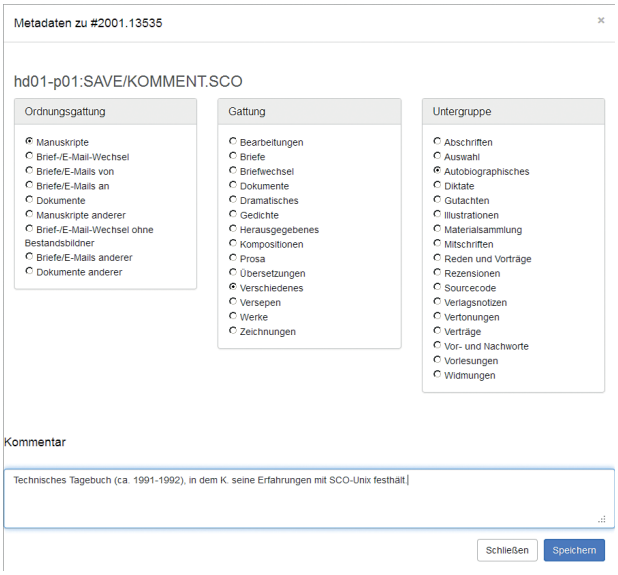

Figure 8. Indexer classification system.

10. CONCLUSION

As should be shown by the article, assigning meaning to digital information is indispensable while facing topics such as long-term access and sustained understanding, research data cycles the preservation needs and the mediation of contextual information over time. Whereas automatized indexing routines enable presorting content, a first result of human interaction is given by a number of grouping routines, which could be established in collaboration with selected archivists and editors. Relating technologically and semantically connected clusters of data with each other, as explained before, provides a good example how far technological skills and semantic knowledge can go hand in hand.

Choosing a less common way of argumentation, our survey tried to explain how far both sides can profit from each other. Whereas parts of the mentioned tasks may be conducted more and more often (and better) by digital tools such as the Indexer, others still require skilled archivists which are familiar with both worlds: the humanities as a field which enables identifying, assigning and documenting meaning in terms of culture, historical, or additional semantic values, and computer science, since technology and their identification get more and more complex.

This leads to at least two points: First, regarding current education and training facilities, a need to cover the cross mix of assigned competences becomes obvious. Especially in Europe, where digital realities in the heritage context have been neglected for too long, certain changes seem to be required. Existing education facilities need to be expanded and at the same time become more attractive to people from different fields of humanities as well as information science. At the same time an image change is required, which deconstructs the cliché of digital culture as nerdy and/or low culture.

The second aspect occurs by facing the big picture of current preservation approaches: Here it seems that (at least) two different types of interest motivate preservation actions today: a) the re-use of data and b) sustainability of authenticity. Whereas in the science sector a strong motivation for (scientific and/or economic) re-use can be observed, ensuring authenticity seems to be the primary aim within the cultural context of memory institutions. Both principles do not necessarily oppose each other. In practice, nevertheless, they can lead to the implementation of varying preservation strategies, parameters and solutions. One example can be found in comparing the way how significant properties or preservation priorities are defined. Archives such as the DLA are positioned at the vertex of these two lines: On the one hand, they are legally bound to preserve the authenticity in the sense of cultural identity. At the same time and at an increasing rate, they are subject to science and the standards of accessibility. However, this intermediate position makes archival involvements in digital preservation actions so interesting. Being routed in both spheres, interest groups of different areas can profit from each other. In this regard, the case of Friedrich Kittler can be seen as paradigmatic: his heritage in humanities will stay only partially comprehensible, without sufficient technical knowledge and vice versa.

11. ACKNOWLEDGEMNTS

Our thanks go to Susanne Holl, Kittler's widow, and to Tania Hron, his colleague, as well as all of his former colleagues. They have all been an invaluable help in technical and especially content-related questions and made us feel the importance of the work we did on Kittler's digital estate.

12. REFERENCES

[1] Berz, Peter and Feigelfeld, Paul. 2013. Source Code als Quelle. Aus der Arbeit mit Friedrich Kittlers Programmierwerk. TU-Lecture, (Berlin, Germany, Aug. 6 2013). URL=https://www.youtube.com/watch?v=kOjGcrj47rk.

[2] Bögeholz, Harald. 2005. Silberpuzzle. Daten von beschädigten CDs und DVDs retten. *c't – Magazin für Computertechnik*, 16 (2005), 78–83. URL=https://shop.heise.de/katalog/silberpuzzle.

[3] Enge, Jürgen, Kramski, Heinz Werner and Lurk, Tabea. 2014. Ordnungsstrukturen von der Floppy zur Festplatte. Zur Vereinnahmung komplexer digitaler Datensammlungen im Archivkontext. In *Beiträge des Workshops "Digitale Langzeitarchivierung" auf der Informatik 2013* (Koblenz, Germany, Sept. 20 2013), 3–13. URN=urn:nbn:de:0008-2014012419.

[4] Holl, Susanne. 2016. Friedrich Kittler's Digital Legacy. Part II. Forthcoming.

[5] Kirschenbaum, Matthew, Farr, Erika L., Kraus, Kari M., Nelson, Naomi, Peters, Catherine Stollar, Redwine, Gabriela. 2009. Digital Materiality: Preserving Access to Computers as Complete Environments. In *iPRES 2009. The Sixth International Conference on Preservation of Digital Objects* (San Francisco 2009), 105–112. URL=http://escholarship.org/uc/item/7d3465vg#page-1.

[6] Kittler, Friedrich. 2011. komment. Computer file. (#1001.10531, text/x-c, 2011-08-18T14:37:46Z,. In: Bestand A:Kittler/DLA Marbach. xd002:/kittler/info [xd, 352.4 KiB]). Internal URL=https://ironmaiden.dla-marbach.de/indexer.final/#id%3A%201001.10531.

[7] Lee, Christopher. 2012. Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision. (International Council on Archives Congress, Brisbane, Australia, August 20–24, 2012). URL=http://ils.unc.edu/callee/ica-2012-lee.pdf.

[8] Lee, Christopher, Woods, Kam, Kirschenbaum, Matthew and Chassanoff, Alexandra. 2013. From Bitstreams to Heritage. Putting Digital Forensics into Practice in Collecting Institutions. BitCurator Project. URL=http://www.bitcurator.net/wp-content/uploads/2013/11/bitstreams-to-heritage.pdf.

[9] NIST. 2016. Introduction to the NSRL. URL=http://www.nsrl.nist.gov/new.html.

[10] Rosenfelder, Andreas and Kittler, Friedrich. 2011. Wir haben nur uns selber, um daraus zu schöpfen (Interview). *Welt am Sonntag,* (Jan 30 2011). URL=http://www.welt.de/print/wams/kultur/article12385926/Wir-haben-nur-uns-selber-um-daraus-zu-schoepfen.html.

[11] Taylor, Hugh A. 1982/3. The Collective Memory. Archives and Libraries. In: *Archivaria 15*. URL= http://journals.sfu.ca/archivar/index.php/archivaria/article/view/10975/11908.

[12] von Bülow, Ulrich. 2003. Rice übt Computer, die Laune wird immer guter! Über das Erschließen digitaler Nachlässe (KOOP-LITERA Symposium, Mattersburg, Österreich, May 08–09, 2003). URL=http://www.onb.ac.at/koop-litera/termine/kooplitera2003/Buelow_2003.pdf.

[13] von Bülow, Ulrich, Kramski, Heinz Werner. 2011. Es füllt sich der Speicher mit köstlicher Habe. Erfahrungen mit digitalen Archivmaterialien im Deutschen Literaturarchiv Marbach. *Neues Erbe: Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung.* Karlsruhe, 141–162. DOI=http://dx.doi.org/10.5445/KSP/1000024230.

[14] Weisbrod, Dirk. 2015. *Die präkustodiale Intervention als Baustein der Langzeitarchivierung digitaler Schriftstellernachlässe*. Doctoral Thesis. Humboldt-Universität zu Berlin. URN=urn:nbn:de:kobv:11-100233595.