9th International Symposium on Electronic Theses and Dissertations
IXe Symposium international sur les thèses et mémoires électroniques
June 7 – 10 Juin, 2006, Quebec City / Québec, Canada

ETD‘s as pilot materials for long-term preservation efforts in kopal

Thomas Wollschläger, German National Library
wollschlaeger@dbf.ddb.de

ABSTRACT

Since summer 2004, the German National Library (GNL) and several partners are conducting a project called kopal, with the objective to establish a technology platform for a long-term preservation repository in Germany.

Electronic theses and dissertations are the first archival objects to be imported into the kopal archive system. They represent the first online publications that have been archived by GNL since 1998, and these over 45.000 documents provide the largest ETD collection in Europe. As a national library, GNL has the task to provide unimpeachable long-term access to these high-value scientific documents. At the same time, their overall consistency, the well-established metadata format and the complete coverage by the stable URN referencing system make them ideal objects for pilot long-term preservation efforts.

The paper includes a general introduction into the kopal concept and its implementation policy. It addresses specific challenges with ETD’s in terms of formats and workflows in the prototypical process. The variety of ETD formats from German universities is being discussed with the focus on how the various formats and their continuing access can be long-term preserved by kopal.

1. CHALLENGES OF THE DIGITAL LONG-TERM PRESERVATION

To maintain a responsible care for their digital collections that have been grown considerably over the past year, universities, libraries and other memory institutions need suitable and consistent archives for the electronic materials. The existing archives in most cases do not, however, fulfil the demands to be a so-called “trustworthy archive”. The considerations of the GNL for its digital long-term archive focussed on the following needs:
- As a basis, the binary data have to be preserved. No existing data carrier is lasting forever or even long enough. The first CD-ROM’s are becoming unreadable already now, and for complex materials (e.g., multimedia applications) the loss of single bits might become fatal. Therefore, the archive should be able to conduct regular bitstream preservation (data carrier migrations). Recent experiences with the non-optimal integrity of certain document servers also indicate that the long-term archive should rely on multiple copies of the archived content in physically separated places.
- The fast technology changes progressional hinder the access to older file formats. Many formats already exist; new ones are being developed, whereas current formats diminish or become obsolete. Additionally, there exist complex dependencies between those formats and various software and hardware environments. The archive should be able to allow constant format migrations (regular conversions) as well as emulations (re-enacting of needed systems).

What are the advantages or disadvantages of these strategies? During migration, older file formats are converted into more recent one’s early enough, as long as they are still readable. That’s being done continuously and means to preserve the integrity and availability of the digital resource in spite a changing environment. There is, however, the risk of possible – but maybe undetected – loss of information (parts) during automatically performed migration routines. The more complex the source format, the more

imminent will become the chance that after long migration chains (over the years) some features may be lost or no longer executable. But migration is, on the other hand, the ideal approach for large amounts of data and will be the most reasonable procedure for the more static data (e.g., text and unmoving pictures). During emulation, a special program (the emulator) tries to re-enact an older system environment onto a present system environment (e.g., the DOS emulation on WinXP platforms). The goal is to be able to execute programs and process data that originally were intended for another, historical system on a recent system. That can be very extensive and presupposes a very exact definition of hardware and software requirements. The advantage of emulations, however, is to be most suitable for complex formats (e.g., multimedia applications) to keep the features of that formats usable as long as possible.

That means, both strategies have advantages and disadvantages; therefore, an archive system whose task is to provide long-term preservation as well as long-term availability of digital resources should be able to use a combination of both strategies. When looking at the format variants of ETD's at the GNL, it becomes clear why that flexibility is of such importance.

2. STRUCTURE OF THE ETD COLLECTION AT GERMAN NATIONAL LIBRARY

The GNL is collecting online doctoral and post-doctoral theses and dissertations since 1997. Their number exceeds 44.500 at present and forms the largest ETD collection in Europe. The annual growth has levelled off at ~ 10.000 ETD's per year, resulting in an accumulated data amount of ~ 350 Gigabytes at present. The ETD's are delivered from the German universities (at present, 83 out of 90 universities are actively delivering their ETD's). All are accessible via the OPAC of GNL, and they are accessible for free and in full-text (except a tiny amount for legal – mostly patent related – reasons). That has led to a widespread access and use of these ETD's, now exceeding 350.000 access cases/month[1].

Corresponding to the scientific value of these documents and their excellent availability for scientific research, the ETD's form the most used and most respected digital collection of GNL. Therefore, it is the major task of GNL to preserve that collection for long-term use. The major challenge for their long-term preservation arises from the fact that the German ETD's are delivered in numerous file formats. From the beginning of the German DissOnline project(s) onwards, the use of innovative file formats has been encouraged over the years; in fact, the use of innovative presentation methods by publishing the theses online has greatly contributed to the great acceptance of ETD's in the German scientific community. These ETD's contain 3-D images & simulations, embedded audio and video files, executables and other file types. However, the first file types are beginning to become hard to access due to the disappearance of suitable viewer applications. It has therefore become imperative to transfer the German ETD's into a suitable long-term archive. On the other hand, if the ETD's at GNL were successfully long-term preserved and their future accessibility ensured, it would have been proven that all the other digital publications that GNL has collected and will collect still, could be subjected to the long-term preservation efforts as well because the highly innovative nature of the ETD's already would have covered all format & content preservation issues. That considerations led to the use of the ETD's as pilot materials for the GNL's long-term preservation efforts within the project "kopal".

3. PRINCIPLES OF THE KOPAL ARCHIVE SYSTEM

[1] For further and more comprehensive information on the German ETD's, consult and/or contact the Co-ordination Agency DissOnline at the German National Library (Homepage: http://www.dissonline.de/ - E-Mail: dissonline@dbf.ddb.de).

"kopal" stands for the Co-operative Development of a Long-Term Digital Information Archive. The project is funded by the German Federal Ministry of Education and Research for a three-year period (to mid 2007; total volume of funding: € 4.2 million). The core of the project is the IBM Digital Information Archiving System (DIAS) solution, developed for the National Library of the Netherlands[2]. The goal of the kopal project is to develop a technological and organizational solution to ensure the long-term availability of electronic publications. After the project financing phase, the resulting archive system will be integrated into the running environment of the participating libraries. Thereby, the transparent integration into existing library systems and the re-usability by memory institutions play a critical role.

In the implementation of the kopal system, international standards for long-term archiving and metadata will be adopted. In this way, both sustainability and the ability to further develop the system are guaranteed. As part of the project, massive amounts of digital materials of all types from two partner organizations, the German National Library and the Goettingen State and University Library, will be deposited. The materials will range from digital documents like ETD's, electronic journals and digitized collections in the form of PDF, TIFF, or TeX files to complex objects like digital videos.

The technical operation of the long-term archive is located at the Gesellschaft fuer wissenschaftliche Datenverarbeitung mbH Goettingen (GWDG). The participation of IBM Germany as a development partner enables the professional customization of the software components. IBM will also provide long-term support.

The archive itself is located at the GWDG site in Goettingen. Objects are being transferred into the system by secure Internet connections. GWDG has a vast experience in safe and secure data hosting and bitstream preservation. GWDG stores the data in two copies at their main site and in two additional copies at a physically different place in Goettingen. Furthermore, from the end of 2006 onwards, a third institution in Munich will receive an additional copy of each object for added security. Thus, the kopal project can fall back on a distributed data storage if needed[3]. Additionally, the logical structure of kopal adheres to several important principles, namely universality, reusability and flexibility.

3.1 Universality

The project aims to build a universally usable archival system in which long-term availability is supported through migration and emulation. There are no limitations within kopal as regards the type of material which can be imported into the archive (text, images, audio, video) and possible data formats. Although the kopal system has a limited total capacity for the project's lifetime, the size of the individual archive object is unlimited. Each of the partners is completely free when it comes to the selection of, and setting of rules for, the import of its collected objects.

3.2 Reusability

Reuse of kopal by other institutions that need long-term archiving is expressly desired. From its inception, the kopal solution has been geared towards a number of different needs. On the one hand, there is the option for a user to have its own "locker" within the system, i.e. secure storage space with data under its own administrative control. This solution is especially appropriate for institutions with a small amount of material to be archived. On the other hand, an institution can reuse the kopal solution by installing the DIAS system itself.

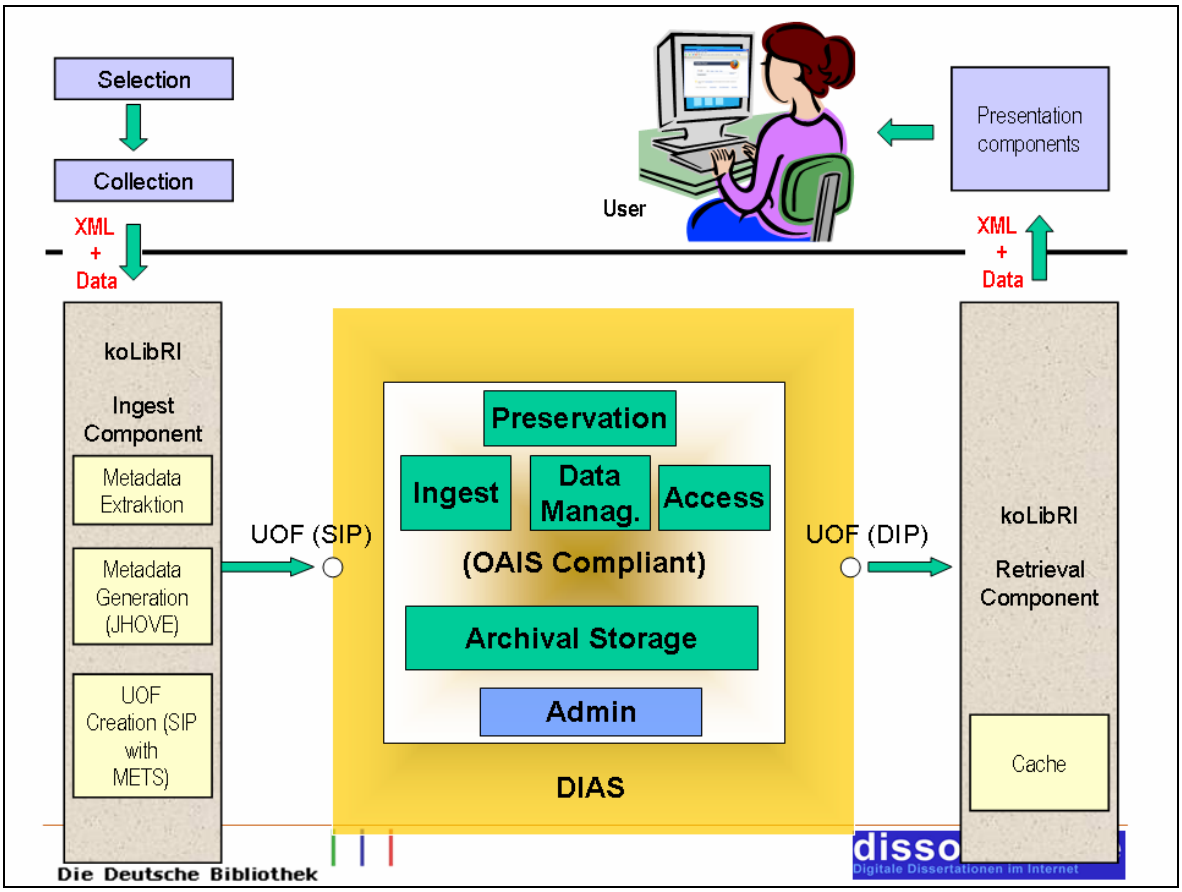
In order to guarantee reusability, established standards are applied. The transfer of objects into a digital archive via standardised formats, paths and interfaces is a key requirement in this context. The kopal

[2] See http://www.kb.nl/dnp/e-depot/dm/dias-en.html
[3] That principle is in some ways comparable to the LOCKSS approach; see http://www.lockss.org/

project has therefore devised the "Universal Object Format" for this purpose, which enables digital objects to be archived with (technical) metadata and exchanged between institutions and systems[4].

3. Flexibility

The DIAS software is also being enhanced with flexible modules. DIAS is based on IBM's standard software components. GNL and Goettingen State and University Library are therefore building software products onto the DIAS core which will be published as the "kopal Library for Retrieval and Ingest" (koLibRI) under an open source licence. These kopal tools primarily support the import of objects into the archive and access to the archived objects. The development of the system is designed to be as open as possible, enabling cooperative use to be extended to all other archiving organisations (libraries, archives and museums) with an interest in reusing the system.



Ill. 1: Outline of the kopal archive system. The inner core refers to the OAIS model. The koLibRI software handles the creation of the archive objects and their retrieval.

4. PRESERVATION PROCESS OF ARCHIVED ETD's BY KOPAL

In effect, the following preservation strategy shall be executable by the kopal system:
- Migrate the object (e.g., a certain ETD) with the precise identifier xxx into the new format yyy.
Or:
- Migrate all objects of (or, more exactly, all files within all objects that contain) the format xxx and/or

[4] See http://kopal.langzeitarchivierung.de/index_objektspezifikation.php.en

that have been ingested before a certain date and/or that are larger than yyy MB into the new format xyz (e.g. from TIFF to PNG).
In fact, the migration of objects with TIFF images into objects with PNG or JPEG2000 images will be the first migration path that shall be executed during the project time of kopal.

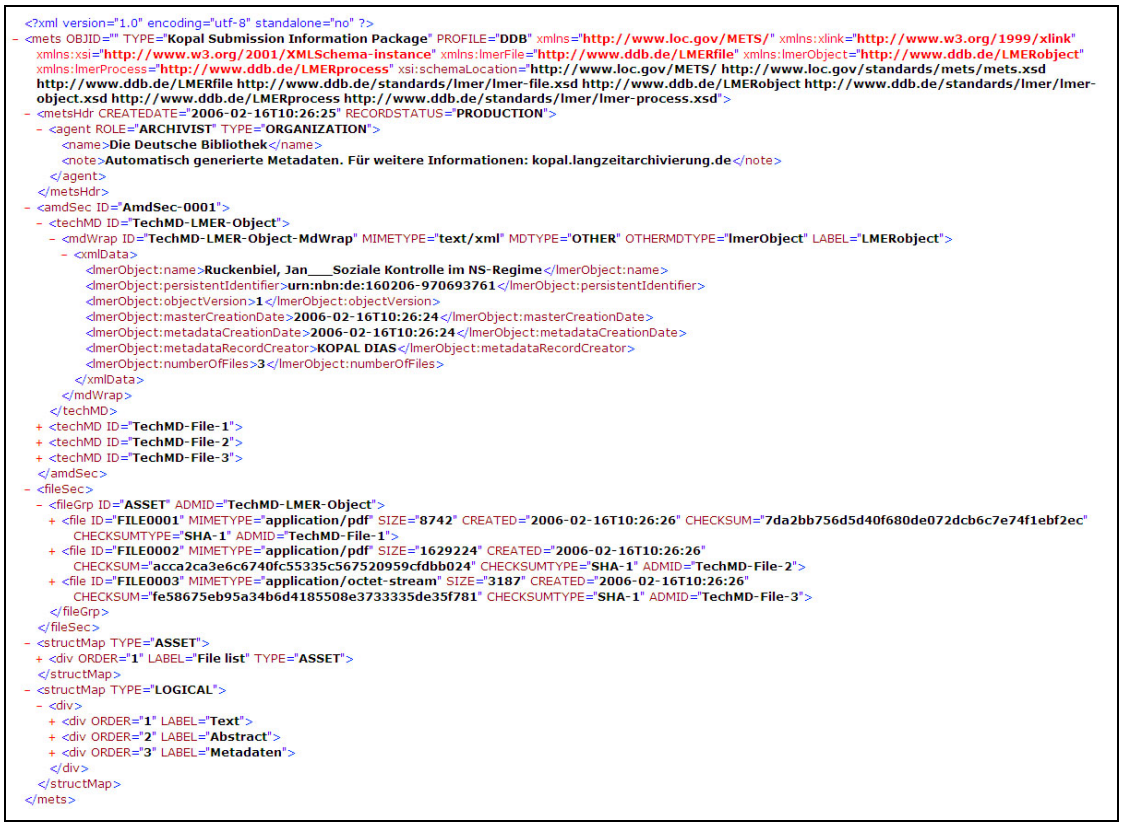
In addition to the migration of objects, the implementation of emulation view paths will take place. Because of the complex structure of emulators and emulation processes, the basis for emulation paths will be implemented as a second step during the course of 2007.

The basis for the execution of migration and/or emulation processes are technical metadata. In kopal, the technical metadata are stored in METS[5] containers that are being packed into one transport unit, together with the actual object, before ingesting into the archive. The METS container includes technical metadata according to the LMER standard (Long-term Preservation Metadata for Electronic Resources)[6].

How are these metadata being created? That process consists of two steps: First, extracting technical metadata as they are delivered with the ETD itself; second, generating additional technical metadata by using appropriate tools.

The first step has an excellent basis in the new metadata format for the German ETD's, the XMetaDiss standard[7]. In the XMetaDiss format, an entire section has been devoted to technical metadata (element <ddb:fileProperties>. It can (and should) contain information on file ID, format, character set, file size, creation application etc. These data are being integrated into the METS file of the archive object. An example for the metadata structure of an archived ETD, including these extracted data from XMetaDiss, is shown in the following picture:

[5] Metadata Encoding & Transmission Standard, see http://www.loc.gov/standards/mets/
[6] http://www.ddb.de/eng/standards/lmer/lmer.htm
[7] See http://www.ddb.de/eng/standards/xmetadiss/xmetadiss.htm



```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
- <mets OBJID="" TYPE="Kopal Submission Information Package" PROFILE="DDB" xmlns="http://www.loc.gov/METS/" xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:lmerFile="http://www.ddb.de/LMERfile" xmlns:lmerObject="http://www.ddb.de/LMERobject"
  xmlns:lmerProcess="http://www.ddb.de/LMERprocess" xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd
  http://www.ddb.de/LMERfile http://www.ddb.de/standards/lmer/lmer-file.xsd http://www.ddb.de/LMERobject http://www.ddb.de/standards/lmer/lmer-
  object.xsd http://www.ddb.de/LMERprocess http://www.ddb.de/standards/lmer/lmer-process.xsd">
 - <metsHdr CREATEDATE="2006-02-16T10:26:25" RECORDSTATUS="PRODUCTION">
  - <agent ROLE="ARCHIVIST" TYPE="ORGANIZATION">
      <name>Die Deutsche Bibliothek</name>
      <note>Automatisch generierte Metadaten. Für weitere Informationen: kopal.langzeitarchivierung.de</note>
    </agent>
  </metsHdr>
 - <amdSec ID="AmdSec-0001">
  - <techMD ID="TechMD-LMER-Object">
   - <mdWrap ID="TechMD-LMER-Object-MdWrap" MIMETYPE="text/xml" MDTYPE="OTHER" OTHERMDTYPE="lmerObject" LABEL="LMERobject">
     - <xmlData>
        <lmerObject:name>Ruckenbiel, Jan___Soziale Kontrolle im NS-Regime</lmerObject:name>
        <lmerObject:persistentIdentifier>urn:nbn:de:160206-970693761</lmerObject:persistentIdentifier>
        <lmerObject:objectVersion>1</lmerObject:objectVersion>
        <lmerObject:masterCreationDate>2006-02-16T10:26:24</lmerObject:masterCreationDate>
        <lmerObject:metadataCreationDate>2006-02-16T10:26:24</lmerObject:metadataCreationDate>
        <lmerObject:metadataRecordCreator>KOPAL DIAS</lmerObject:metadataRecordCreator>
        <lmerObject:numberOfFiles>3</lmerObject:numberOfFiles>
       </xmlData>
     </mdWrap>
    </techMD>
  + <techMD ID="TechMD-File-1">
  + <techMD ID="TechMD-File-2">
  + <techMD ID="TechMD-File-3">
  </amdSec>
 - <fileSec>
  - <fileGrp ID="ASSET" ADMID="TechMD-LMER-Object">
   + <file ID="FILE0001" MIMETYPE="application/pdf" SIZE="8742" CREATED="2006-02-16T10:26:26" CHECKSUM="7da2bb756d5d40f680de072dcb6c7e74f1ebf2ec"
      CHECKSUMTYPE="SHA-1" ADMID="TechMD-File-1">
   + <file ID="FILE0002" MIMETYPE="application/pdf" SIZE="1629224" CREATED="2006-02-16T10:26:26"
      CHECKSUM="acca2ca3e6c6740fc55335c567520959cfdbb024" CHECKSUMTYPE="SHA-1" ADMID="TechMD-File-2">
   + <file ID="FILE0003" MIMETYPE="application/octet-stream" SIZE="3187" CREATED="2006-02-16T10:26:26"
      CHECKSUM="fe58675eb95a34b6d4185508e3733335de35f781" CHECKSUMTYPE="SHA-1" ADMID="TechMD-File-3">
    </fileGrp>
  </fileSec>
 - <structMap TYPE="ASSET">
  + <div ORDER="1" LABEL="File list" TYPE="ASSET">
  </structMap>
 - <structMap TYPE="LOGICAL">
  - <div>
   + <div ORDER="1" LABEL="Text">
   + <div ORDER="2" LABEL="Abstract">
   + <div ORDER="3" LABEL="Metadaten">
    </div>
  </structMap>
</mets>
```

Ill. 2: Example of the mets.xml file of a kopal object with the technical metadata of the archive object. It represents an ETD, consisting of a main document ("Text"), an abstract and bibliographic metadata. The appropriate tags in the File Section (<fileSec>) contain information extracted from the XMetaDiss metadata.

In addition to the extracted metadata, the tool JHOVE[8] is being used to generate additional technical metadata for each file of the ETD to be archived. In the above picture, the JHOVE output would be stored in the "techMD" sections for each file of the ETD. We are confident that this structure will allow the mentioned preservation strategies.

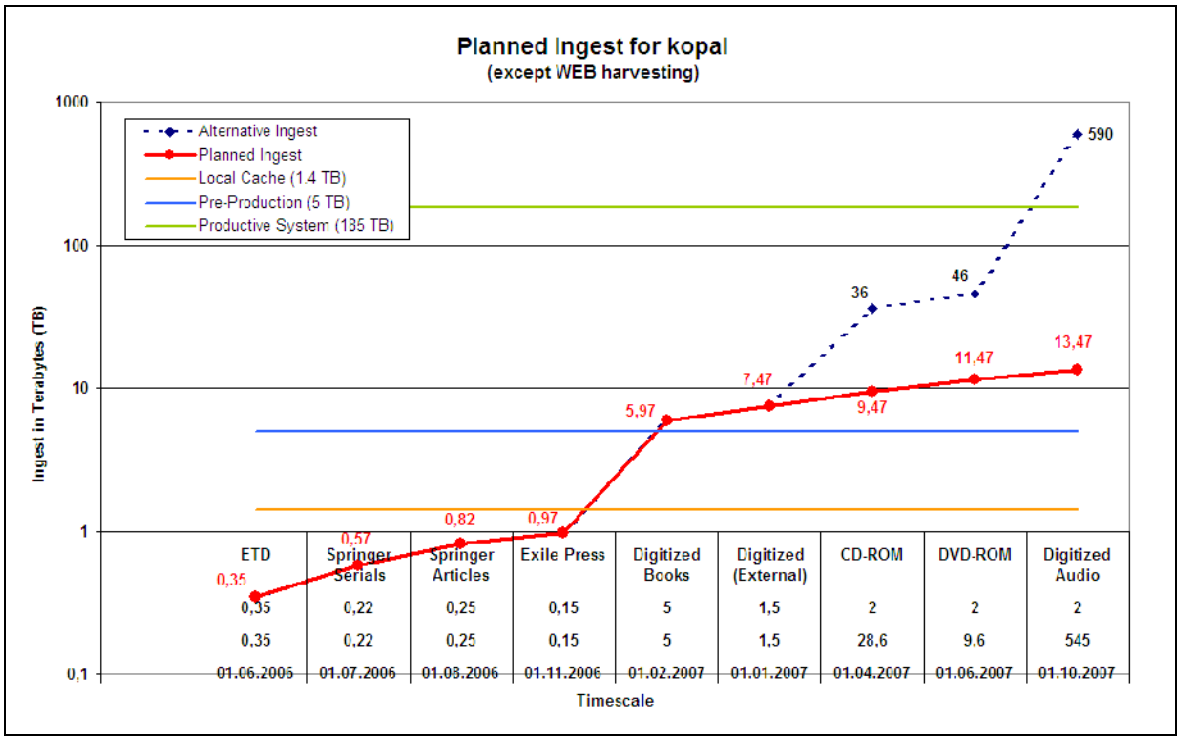
4. MATERIALS FOR PRESERVATION AFTER ETD PILOT

As mentioned before, the successful ingest and exemplary migration of the ETD collection of GNL is the precondition for the planned ingest of all the other digital materials that GNL and Goettingen University possess. In case of GNL, that materials cumulate to quite a considerable amount of data:

- Electronic journals & serials, data amount: ~ 300 GB
- CD-ROM images, number: ~ 50.000 to 100.000, data amount: ~ 28.000 to 56.000 GB
- Digitized materials:
  - Exil Press Digital (from GNL): ~ 150 GB
  - External digital collections: ~ 1.500 GB
  - Digitised books from the German Book & Scripture Museum (GNL): ~ 5.000 GB (for starters)
  - Born-digital and digitised audio from the German Music Archive (GNL): ~ 544.000 GB

[8] JSTOR/Harvard Object Validation Environment (in short: JHOVE) version 1.0 (release of 05-26-2005) with some bugfixes which will be also contained in the upcoming maintenance release; see http://hul.harvard.edu/jhove/

It should be mentioned that the carrier-based digital collections have been digitized not nearly complete by now; in fact, that process will take several years. But already now, the kopal system has been design to be able to store more than the readily available data amount. An impression of the relationship between the potential materials that will be ingested into kopal, and the planned capacity for the project phase, is shown in the next picture:



Ill. 3: Planned data ingest for the kopal system with ETD's as the first content. The red line indicates the realistic amount of data to ingest until end-2007 (logarithmic graph).

5. NEXT STEPS & VISION

A major part of the project work will be the integration of the archive system in the existing workflows of the library and the eventual creation of new workflows that incorporate the preservation strategy for the ETD's and other materials right from the delivery by the publishers. Since kopal does not provide an end-user access or a direct ingest from the outside (as can be seen in Ill. 1, kopal supports and/or provides interfaces to the collection/acquisition side and to the user services), these components – the user interface and the placing of the materials to be ingested at the disposal of the kopal tools – have to be provided by the appropriate departments of the GNL. Of course, these departments, the IT department and the kopal team are working closely together to set up suitable interfaces and workflows.

One example might illustrate a resulting task. In face of rising data amounts and large single objects (e.g. digitised DVD-ROM images with more than 8 GB each), it is very important to guarantee a sufficient performance of the system. When an end-user clicks on a link within the search result display in the OPAC, the retrieval of the archived ETD from the kopal system starts. But, in case of those very large objects, it might take several minutes to get the object via the Net onto the screen. To prevent users from multiple clicking or simply walking away, fast Internet connections have to be provided for, and suitable access systems have to be implemented that facilitate an appropriate user support (e.g., including messages on the needed retrieval time).

The other major working package includes the implementation of a functioning Preservation Planning mechanism. Since the reference to precise file formats is an essential of any migration or emulation mechanism, the best support for the kopal – and any other long-term preservation effort – would be the successful setup of a functioning international File Format Registry. Several international efforts, in which GNL is participating, are currently under way to cover that area. While that efforts are under way, kopal has to prove the performant migration of large data amounts as well as the successful implementation of emulation mechanisms until the end of 2007.

And, finally, GNL and its partners will broaden their information, support & encouragement of ETD producers (including authors and universities) towards a more well-founded format & preservation awareness. Projects as "DissOnline Tutor"[9] with the aim of improving the technical quality and develop & procure tools to create and control long-term archiveable ETD's are an example of these efforts. We also want to encourage other memory institutions to adapt to long-term preservation solutions in order to protect, preserve and make available the digital cultural heritage of our times for future generations.

On the author:

Dr. Thomas Wollschläger is a digital preservation officer of the kopal project, and works for the Department of Information Technology at the German National Library in Frankfurt/Main. Within the context of long-term preservation of digital data and especially within kopal, he is responsible for the integration of the archive system in the workflows for electronic materials in the library, the co-ordination of the departments and of external partners as well as for supporting the project controlling.

[9] http://www.dissonline.de/Tutor/ (available only in German)