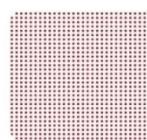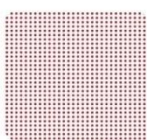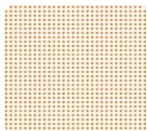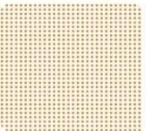![ZB MED Wissen für Mensch & Umwelt logo]

Digital Preservation of Biomedical Documents: State of the Art

Ulrich Korwitz, Michelle Lindlar - German National Library of Medicine (ZB MED)

Brussels, 5.7.2012

German National Library of Medicine

- Largest European Library in the fields of Medicine, Health, Nutrition, Environment and Agriculture
- Located in Cologne and Bonn

German National Library of Medicine

- Hugh collection of monographs, journals and multimedia material, also gray material
- Responsible for archiving to guarantuee a perpetual access to all holdings
- Part of the German National Library System
- Responsible for the cultural/scientific heritage in Germany in its subject fields

→ **Long term preservation as a new focus in ZB MED**

Long term preservation

Long term preservation is the collection, the long range storage and the care for the **continuous availability** of information.

„ ‚Long term' in the context of the preservation of digital resources does not mean giving a guarantee for five or fifty years, but the **development of a strategy** coping with the everlasting change in the information industry."

(From: H. Neuroth et al., „nestor Handbuch" Version 2.3)

„*The NASA-Effect*"

„Pioneer" Mission 1979:
Data transmitted from Pioneer spacecraft were stored by
NASA on four different data carriers (9-track magnetic tapes,
7-track magnetic tapes, punch tapes and punch cards).

In 1994, all the data were not readable anymore because of
a lack of suitable reading devices for these data carriers

Risks

Hardware to read the data carrier is not available anymore

Data carrier nonreadible because of wearing/erosion or environmental damage

Software is not running on the current operating system

Risks– Data Carriers

Data Carrier	Mean Lifecycle
Acid free paper	Several hundred years
CD-R/RW (burned)	5-10 years for CD-R, less for CD-RW (no reliable studies available)
CD-ROM (stamped)	10-80 years (depending from usage and storage conditions)
Floppy disks	10-30 years
Hard disks	2-10 years (if in continuous use)
Magnetic tapes	Min. 30 years
USB-Sticks	3-10 years (depending from usage)

Risks – Technological Progress

The technological progress is a risk for the availability of data:

- Software to interpret the file format does not exist any more
- Software exists but cannot be executed on the operating system
- Software runs but the format does not match user expectations

Digital Holdings of ZB MED

„Offline"	„Online"
Thesis of Universities (CD-R, CD-RW)	ElliNET Repository
Supplementary materials to journals (CD-ROM, DVD, Disks)	gms Open Access Publikations
Supplementary materials to monographs (CD-ROM, DVD, Disks)	Outputs from retro-digitisation
Backups of National licences (HDD, DVD)	eyeMoviePedia
Supplementary material to thesis (CD-R, CD-RW, Disketten)	CoCoMac
AV-Materials (DVD, LD)	In the future: Virtual Microscopy

Digital Holdings

- Heterogeneous Data
- High growth expected especially in the field of online content
- Digital research data (primary data) play a more important role
- Strategy and actions have to be worked out and checked/updated yearly

„Offline" Holdings

- Strategy:
  - Preservation of the data (bitstream-level) by replication
  - Preservation of information (format-level) by migration oder emulation
- Action:
  - Survey and sample analysis of data carriers in the holdings (CALIPR)
  - Priorisation
  - Ingest into a digital preservation system

CALIPR

PRESERVATION PLANNING

CALIFORNIA PRESERVATION PROGRAM

- CALIPR home
- Introduction

ACTIONS
- Login for registered users
- New user registration

TOOLS
- CALIPR help

CALIPR

A collection needs assessment instrument for preservation planning

CALIPR was developed at the UC Berkeley Library for the California Preservation Program with support from the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. The opinions expressed herein do not necessarily reflect the position or policy of these organizations and no official endorsement should be inferred.

Results of a sample analysis

- Results of an autopsy: 89% „good", 10% „medium", 1% „poor" condition
- Results of tests for readability: 3% non readable (4 CDs; autopsy said: „good condition")
- High variation in relation to CD-R manufactorers (30 brands)
- 28 different data formats, with PDF = 90%

„Online" holdings

- Strategy:
  - Preservation of publications (document repository) and of primäry data on bitstream-level by replication und refreshing
  - Preservation of information (format-level) by migration oder emulation - presentation systems not preserved
- Action:
  - Process analysis
  - Priorisation
  - Ingest into a digital preservation system

Example of a Process analysis

„ElliNET. Medicine. Health.“

- Established workflow for the acquisition of medical thesis and other electronic gray literature

- Homogeneous data (pdf, zip, xml)

- Technical platform: DigiTool, hosted on a realiable server outside ZB MED

- Data volume: 1108 Objekte, 2 GB

Results of the process analysis

- Complete process with defined responsibilities
- Metadata are stored with original data as Dublin Core in xml
- Publication contract between authors and ZB MED allows migration
- Risk: Updates for DigiTool are not always made available on time by host

GOPORTIS
Leibniz–Bibliotheksverbund Forschungsinformation

STARTSEITE   ÜBER GOPORTIS   KOMPETENZEN   KOOPERATIONEN   AKTUELLES   PRESSE

Goportis ist der Verbund der drei Deutschen Zentralen Fachbibliotheken.

GOPORTIS

Bildnachweise

Wir bieten forschungsbasierte Services, um wissenschaftliche Arbeitsprozesse zu unterstützen.

Aktuelles

❯ „Exzellente Forschung braucht exzellente Forschungsunterstützung. Dazu setzen die drei großen… ...mehr

❯ Das Thema Forschungsdaten stand im Mittelpunkt eines Vortragsblocks von Goportis auf dem 101… ...mehr

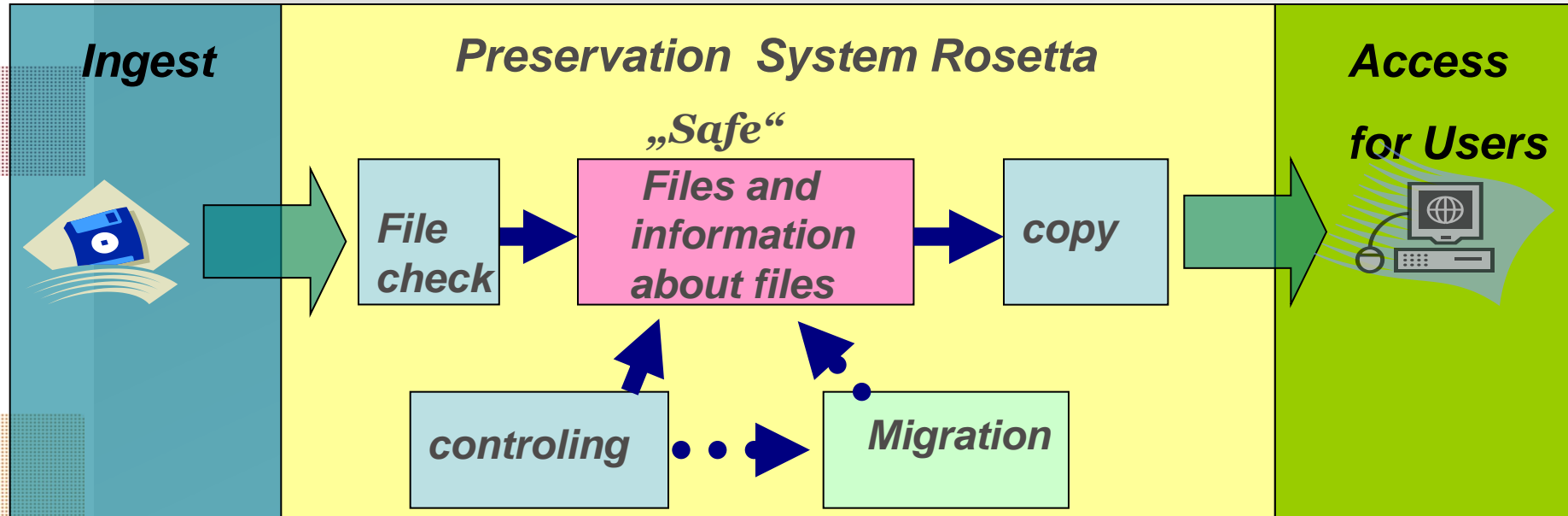❯ Mit einer strategischen Neuausrichtung wird sich Goportis auf dem

Meinungen

❯ „Exzellente Forschung braucht exzellente Forschungsunterstützung. Dazu setzen die drei großen Spezialbibliotheken in Deutschland bei der Entwicklung ihrer Dienstleistungen konsequent auf anwendungsorientierte Forschung sowie international renommierte Kooperationspartner. Dieser Weg stärkt nicht nur die Leibniz-Gemeinschaft, sondern den gesamten Wissenschaftsstandort
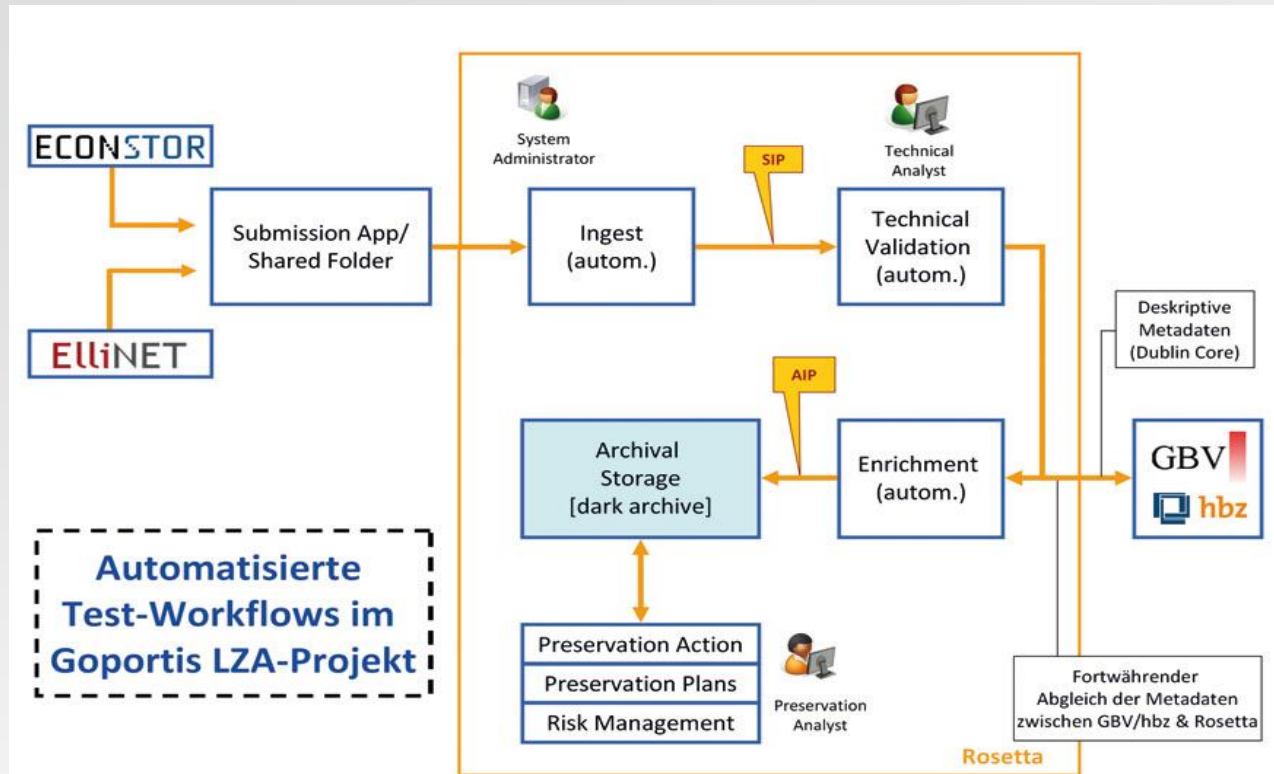
Criteria for the decision in favor of Rosetta

- Multi-client capability, permission management capability
- Open interfaces via API, SDK
- Open standards: OAIS, PREMIS, METS, DC, etc.
- Open source tools: JHOVE, DROID, NLNZ metadata Extractor, etc.
- All necessary modules: Ingest, Preservation Planning, Preservation Action, Storage, Access
- Out-of-the-box system for these modules
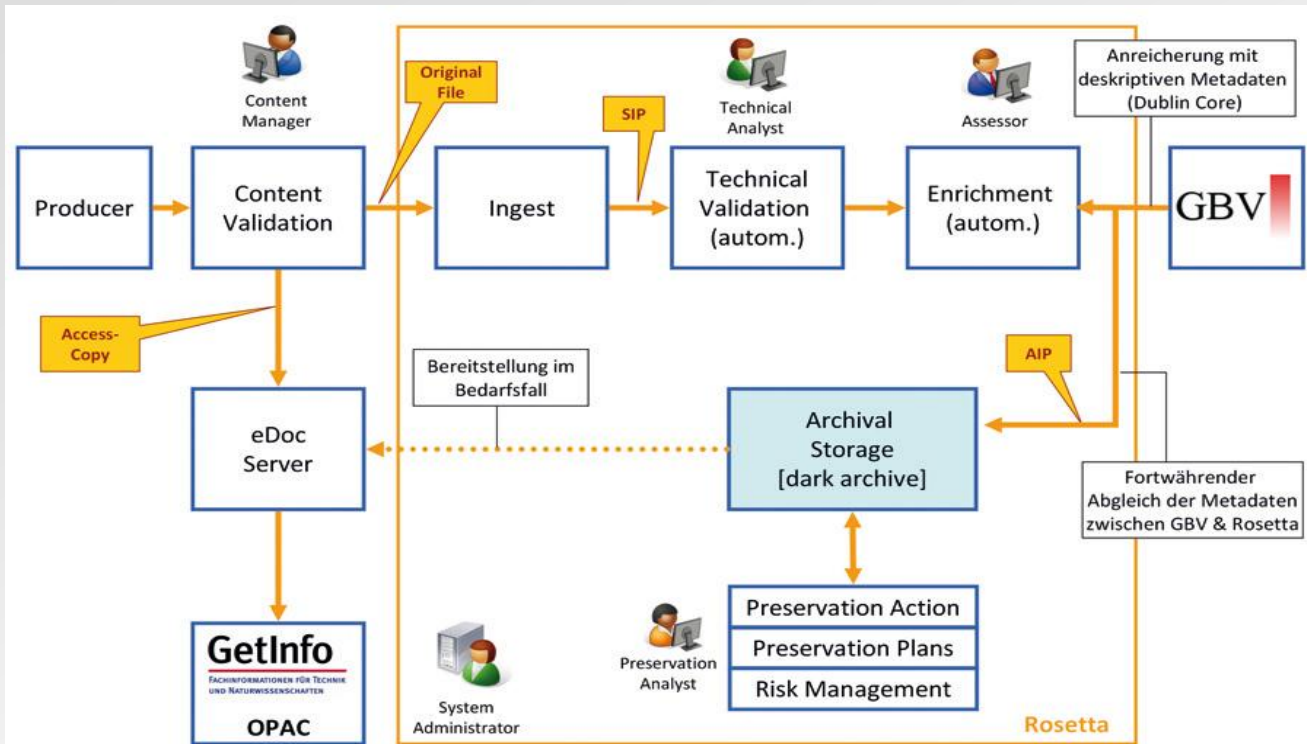- Configuration and administration of all modules possible; scalability

Stucture of the Digital Preservation System

Ingest

Preservation System Rosetta

„Safe"

File check

Files and information about files

copy

controling

Migration

Access for Users

Automated Workflows

Manual Workflows

Co-operation

Intentions:

- Active shaping of national and international strategies
- Introduction and evolution of open standards and tools
- Co-operativ exchange of experiences
- Monitoring of new developments

Co-operation

Responsibilities

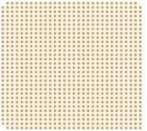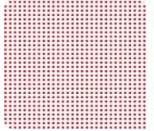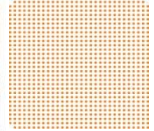- Who should care for digital preservation?
- What system should be selected?
- Who is monitoring the technical progress?
- Who pays the costs?
- What outcome is desirable?

Thank you !

More information: Korwitz@zbmed.de