Digitale Langzeitarchivierung – können wir uns das leisten?

Gerhard Schneider

Rechenzentrum Universität Freiburg

direktor@rz.uni-freiburg.de

UNI FREIBURG

Fakten

- Immer mehr Dokumente liegen „nur“ noch elektronisch vor
  - Oder werden (mühsam) digitalisiert
- „Aufheben“ erfordert etwas mehr Nachdenken als das Lagern von Papier
  - Papier ist normalerweise nicht beleidigt, wenn sich 200 Jahre lang niemand darum kümmert
  - Elektronische Speichermedien veralten technisch schneller als dass sie vergammeln
  - „wiederfinden“ – was bedeuten 0 und 1 konkret?
    - Auch das Wissen über Hieroglyphen ging mal verloren
- Erste Strategiepapiere … ☺ ☹

Fakten

- Elektronische Zeitschriften verschlimmern die Situation
  - Zugriffsrechte nur auf Zeit ??
  - Open access journals mit unklarer Betreiberlage
- Mit dem „Tod“ des Betreibers verschwinden die Daten – selbst auch wenn sie allen gehört hätten
  - Und Amazon hat schon mal ein Buch vom *kindle* gelöscht, obwohl es bezahlt war
  - Zur Abgrenzung des Marktes nutzen die Hersteller propietäre Formate, die nur mit ihrer Hardware lesbar sind
    - Aktuelle Tendenz im ebook-Markt!

Urheberrecht

- Mit Hilfe des Urheberrechts werden die proprietären Welten verteidigt
  - Licence to print money
  - Wer hindert einen Hersteller daran, das alte Format für obsolet zu erklären, um die alt bekannten, aber rechtlich wohl geschützten Inhalte den Kunden nochmals zu verkaufen?
    - Schallplatte -> CD -> Kopierschutz -> blueray
    - Offene Format / Offene Technik (wie PC) schon auf dem Rückzug?
    - ACTA … ?!?!

Bibliotheken ??

- Welche Aufgabe hat eine Bibliothek?
  - Aufheben und wiederfinden
- Hat diese Aufgabe in der Zukunft dann noch Bestand?
  - Aufheben, wiederfinden, als nicht nutzbar erkennen, wegwerfen ????
- Aufgeben?

- Noch mal von vorn…

Der einfache Teil

- Daten – **wenn** sie im Sinne von „Besitz“ vorliegen – liegen auf Datenträgern
- Und können in den Rechenzentren von einem Datenträger auf die nächste Generation umkopiert werden
  - Ohne Qualitätsverlust!
  - Qualitätssicherung des Kopierverfahrens kein Problem
- Aber: die Zeit ist ein Faktor!
  - Im kleinen Freiburger Uni-RZ: Bandroboter mit (demnächst) 3 PetaByte Speicherkapazität

Der einfache Teil (2)

- Freiburg:
  - 16 Laufwerke TS1120 mit je 100 Mbyte/sec
  - Das Auslesen aller Daten dauert also im **günstigsten Fall** (reine Theorie): 22 Tage
  - Neue TS1140-Laufwerke: 2,5 x schneller
  - Verzögerungen durch zwischengeschaltete Hardware, Übertragungskanäle, etc
- Die reine Datenmigration auf neue Technologie wird wohl 6-12 Monate dauern
  - Es muss also **frühzeitig** vor Auslaufen der alten Technologie begonnen werden
  - Das widerspricht dem öffentlichen Finanzgebaren.
- Anderer Vorstellungswelt:
  - das Füllen eines 16GB USB-Stick dauert etliche Stunden

Der einfache Teil (3)

- Aus RZ-Sicht ist die Langzeitarchivierung eine Art „Lebensversicherung“
  - Keiner kann‘s besser und es ist Kerngeschäft
  - Wenn man das RZ schließt, sind die Daten weg
    - Und damit das kulturelle Gedächtnis
    - Und es ändert sich nicht, wenn sich das RZ einer Bibliothek daran versucht: mind. genauso teuer ☺
- Erkenntnis: digitale Langzeitarchivierung erfordert permanente Investitionen
  - Naja, Bücher wollen auch irgendwie gepflegt werden → Klimaanlage, etc
- **Gute Nachricht:** die notwendige Technologie ist nicht übermäßig teuer - und automatisierbar

Was erwartet die Bibliotheken?

- Wer in
  - Digitalisierung = eingescannte Bilder/Dokumente
  - PDF-Dokumenten
  - Statischen Webseiten (html zum Abernten)

  denkt, wiegt sich in falscher Sicherheit
- und trifft falsche Zukunftsentscheidungen ☹
- Vorahnung / Vorgeschmack:
  - Durchforsten Sie mal Ihre Bestände nach nachgewiesenen Computerspielen, frühen e-learning Umgebungen, etc
  - Also Daten, die eine Ablaufumgebung benötigen, um dargestellt zu werden
  - Was machen Sie damit? Wegwerfen?
    - Was sagen die Soziologen dazu, die die Entstehung von virtuellen Umgebungen = Computerspielen erforschen wollen?

Was erwartet die Bibliotheken (2) ?

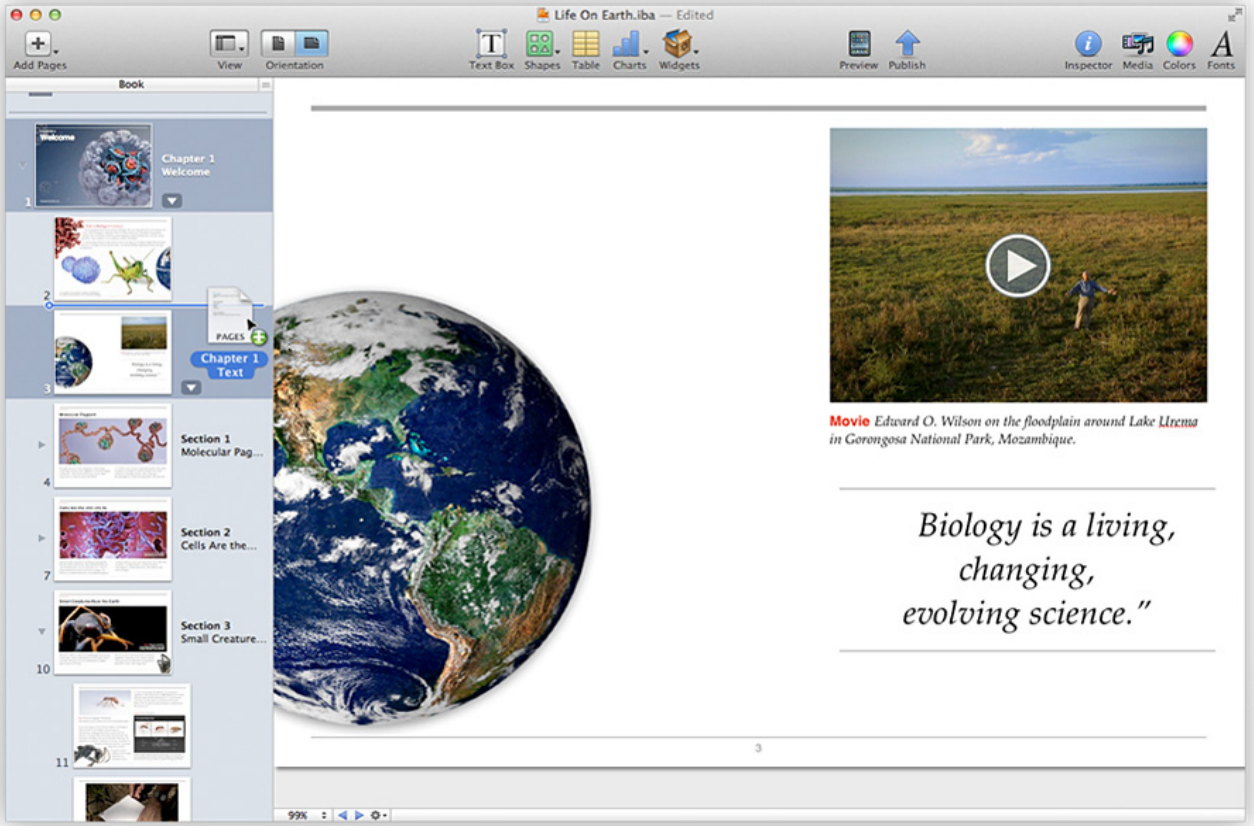
- Digitale Daten sind im Zeitalter der Vernetzung nicht mehr notwendig statisch
  - Beispiel: ein Lehrbuch, das Wirtschaftsverständnis über Börsenspiele mit aktuellen(!) Daten vermittelt
  - Dabei hilft nicht, wenn manche Professoren dies für didaktischen Unsinn halten
    - Es kommt trotzdem
  - Beispiel: ein Lehrbuch, in dem Verständnis über tatsächliche Simulationsrechnungen vermittelt wird
    - „Energiewende“ über sich ändernden Energiemix unter Nebenbedingungen (wie Speicherkapazitäten)
- So was gibt’s noch nicht?

Apple ….

- Baut ein Ökosystem:
  - Software Apple iBooks author …
    Zitat aus http://www.apple.com/de/ibooks-author/
    iBooks Author. Tolle Multi-Touch Bücher für das iPad erstellen und veröffentlichen
  - iPad als ebook-Reader (online via WLAN, UMTS)
    - Endlich „ubiquitous computing“, „always on“
    - Touch-interaction
      - Nicht nur blättern, sondern Auslösen von Aktionen
    - Multimedia (Video, Ton, Bildfolgen, Simulationen)
      - Integriert auf **einer** Seite, inline
      - Inhalte werden online nachgeladen
        - und sind nur in einem Zeitraum verfügbar (?)
- Zum Vergleich: Schulbücher mit 15 Jahre alten Daten……

Apple …. (2)

- *Mit Galerien, Videos, interaktiven Diagrammen, 3D Objekten und mehr lassen diese Bücher ihre Inhalte auf eine Weise lebendig werden, wie es gedruckte Bücher niemals könnten.*



- Ignorieren??
  - Vielleicht/vermutlich die Zukunft!
  - 10 Seiten Bleiwüste lesen, um die Zellteilung zu verstehen – oder einen guten Kurzfilm ansehen?

Herausforderung für Bibliotheken ?

- Und wie hebt man sowas auf?
  - Den iPad in eine Schatulle legen?
    - Quasi als Buch behandeln?
    - Und was ist mit den Online-Inhalten?
    - Ist der archivierte iPad in 10 Jahren leer?
      - Akku ist dann eh kaputt und nicht wechselbar
- Einzelfall?
  - DFG wünscht Aufbewahrung der Forschungsdaten für 10 Jahre
    - *Genauer: die Wissenschaftler und Wissenschaftler wollen dies und fordern dies via DFG*
  - Naiv: Forschungsdaten fallen wie Wetterdaten an
  - Real: elektronische Laborbücher sorgen für Ordnung im Labor

Herausforderungen …

- Elektronische Laborbücher sind proprietäre Umgebungen mit proprietären Speicherformaten
  - Weil die Technik neue Formate benötigt und noch keine Normierung erfolgt ist
    - Vielleicht nie erfolgen kann ☹
  - Zum Auslesen und Darstellen der Daten benötigt man eine Laufzeitumgebung
    - die in 10 Jahren 10 Jahre alt sein wird ☹
      - Und vielleicht nicht mehr existiert
      - Vielleicht gibt's in 10 Jahren kein Windows mehr, nur iOS oder Android (oder genau anders herum)
  - Selbst wenn man die Daten noch auslesen könnte – wer kann die Software dann noch sinnvoll bedienen?
    - Software-Bedienung ist eine zeitabhängige Kulturkomponente

Herausforderungen

- DFG rüstet nach: Datenmanagement-Konzept
  - Schon besser, aber immer noch zu kurz
  - Verhindert immerhin das „Brennen auf DVD, Lagerung in der Schublade“
- Digitale Langzeitarchivierung beinhaltet also auch:
  - Langfristige Verfügbarkeit der Daten und des gesamten Umfelds
  - Sonst können wir eines Tages nicht mehr nachvollziehen, wie wir einst gelernt haben
    - Das klassische Schulbuch zum Nachlesen gibt’s dann ja nicht mehr
    - Schlimm? Vielleicht nicht.
    - Aber wenn selbst die Basisdaten für den Erkenntnis-gewinn verschwunden sind? *KTvG hätt‘s gefreut*

Herausforderungen

- Langsam wird klar:
  durch die Unzulänglichkeiten des Speichermediums Papier war man gezwungen, so zu abstrahieren, dass alles zweidimensional statisch (kompliziert für: *gesetzt und gedruckt*) abgelegt werden kann.
- Dennoch: Naturwissenschaften hatten schon immer Probleme, komplexe Experimente sauber „auf Papier“ darzulegen
  - Ok, wenn experimentelles setup leicht zu beschreiben und nachvollziehbar
  - CERN: die Messwerte hängen doch auch vom gebauten Beschleuniger ab, egal wie viele Daten man protokolliert
    - Oder man kann später beweisen, dass das setup fehlerhaft war

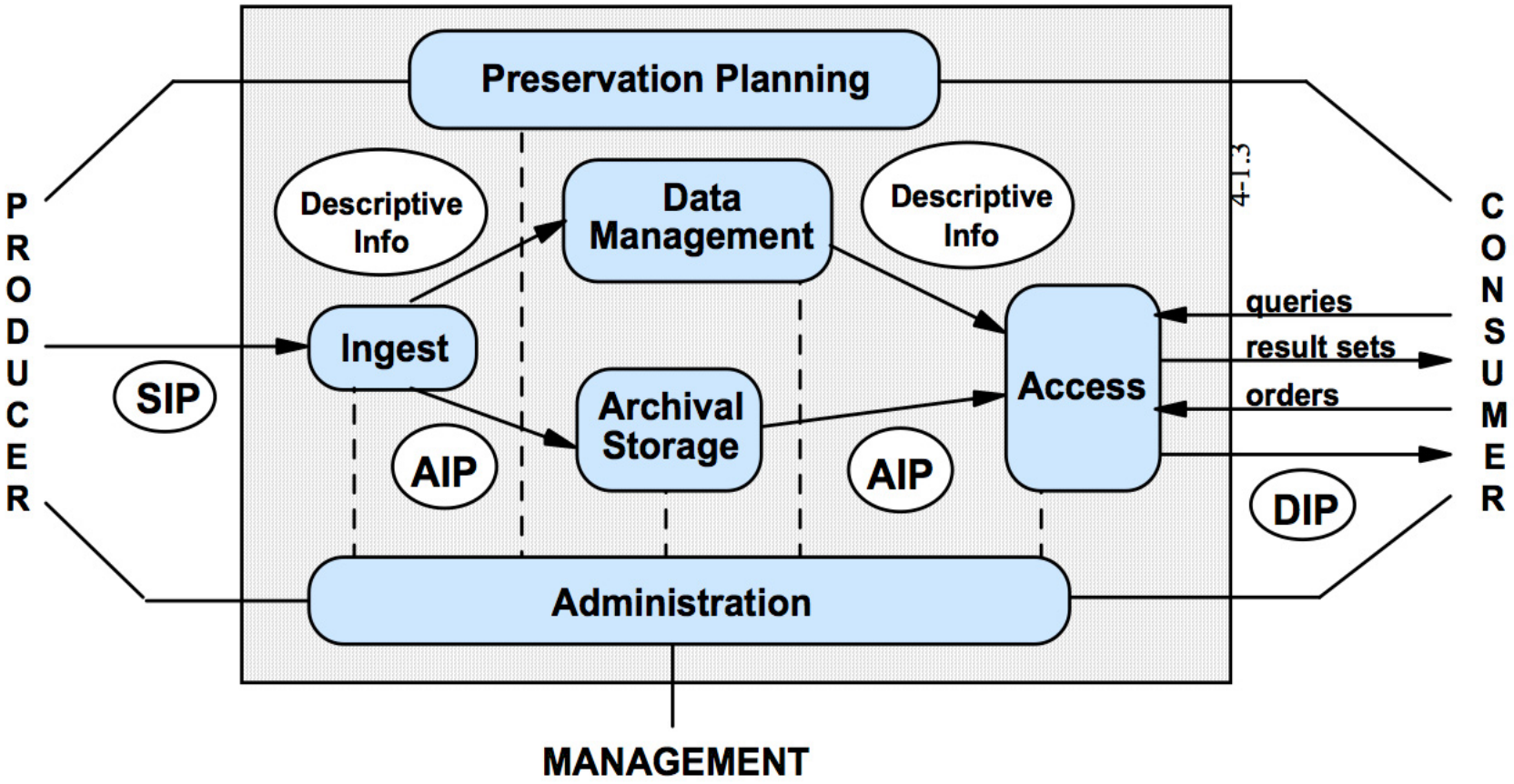
Herausforderungen

- Die digitale Welt ermöglicht umfassendere Wissensvermittlung
  - Darstellung zur Laufzeit
- Die Archivierung der Daten ist komplexer
  - Archivierung der Laufzeitumgebung
  - Archivierung der Online-Umgebung
  - *Archivierung des aktuellen Zustands der Welt* ☺
    - Selbst Facebook archiviert nur das, was es zuläßt
- Die Abstraktion von den Möglichkeiten auf das „Stemmbare“ hat noch nicht eingesetzt.
- Was tun?
  - Fassungslos zusehen?

Lösungsansätze in Teilbereichen

- „Manche“ Daten benötigen zur Darstellung Software
  - In einer speziellen Version
    - Winword 2.0 …
  - die nur in einer speziellen Umgebung lauffähig ist
    - Win95, NT,….
  - die spezielle Hardware und Treiber benötigt
    - Kraftwerkssteuerung …, Apple Newton, PalmPilot
  - die also nur in der Vergangenheit funktionierte
- Ansatz Migration:
  - Daten werden mit der letzten Software-Version, die diese Daten noch darstellen kann, ausgelesen und in einem neuen Format gespeichert
    - Darstellungstreue?
    - Rechnet Excel dann noch richtig?

Long-Term Archival Model (OAIS)

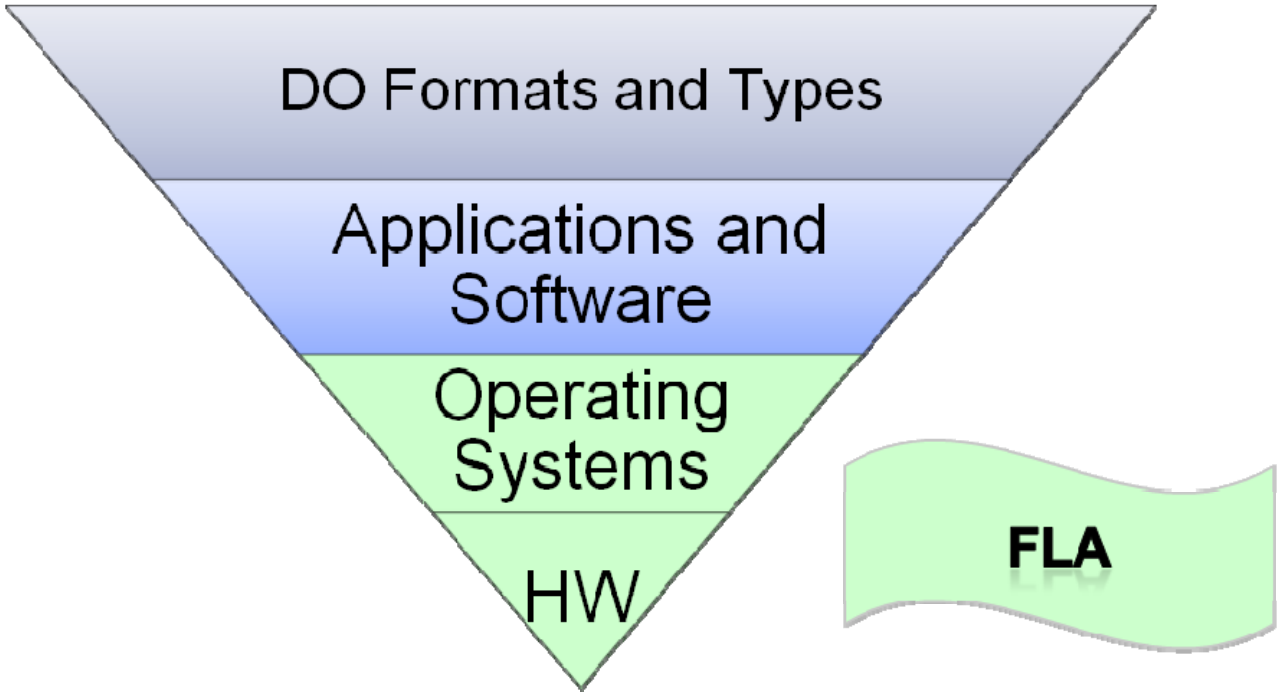


- Rechnung ohne den Wirt: Abhängigkeit von der Umgebung nicht bedacht
- Damit hält das Konzept nicht lange ☹

Lösungsansätze

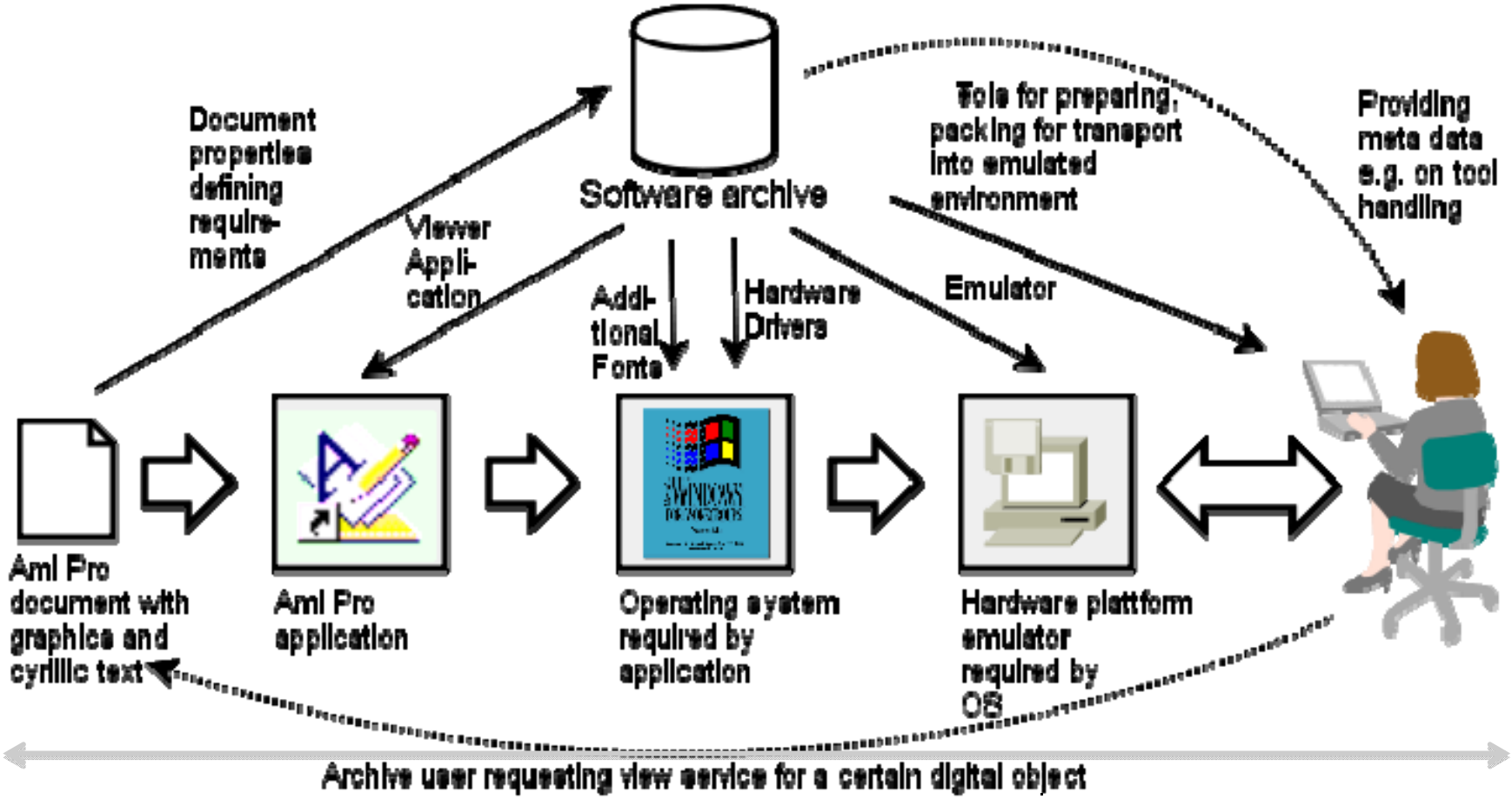
- Funktionale Langzeitarchvierung
  - Archivierung der gesamten Laufzeitumgebung
    - Als virtualisierte Welt
      - Hardware → Software
    - Gesamte Umgebung wird zu einer „Datensammlung“
      - Zip-Datei, enthält auch den Rechner
    - Die berechtigte Hoffnung besteht, dass die virtualisierte Welt auch in Zukunft ausführbar sein wird
  - Archivierung von Bedienungs-Szenarien
    - Als Screen-Mitschnitte echter Szenarien
- Was braucht man denn so an Metadaten?

bwFLA – Funktionaler Ansatz



- Funktionaler Ansatz
  - -Bewahrung von Systemumgebungen
  - -Emulation als Werkzeug
    - Zugang zu DO, ie. rendering, Nutzung und Migration
  - → Preservation Planning / Preservation Action
    Fokus auf eine „kleine Komponentenmenge“

Viewpaths & SW-Archive



View-Path eines Lotus AmiPro Dokuments

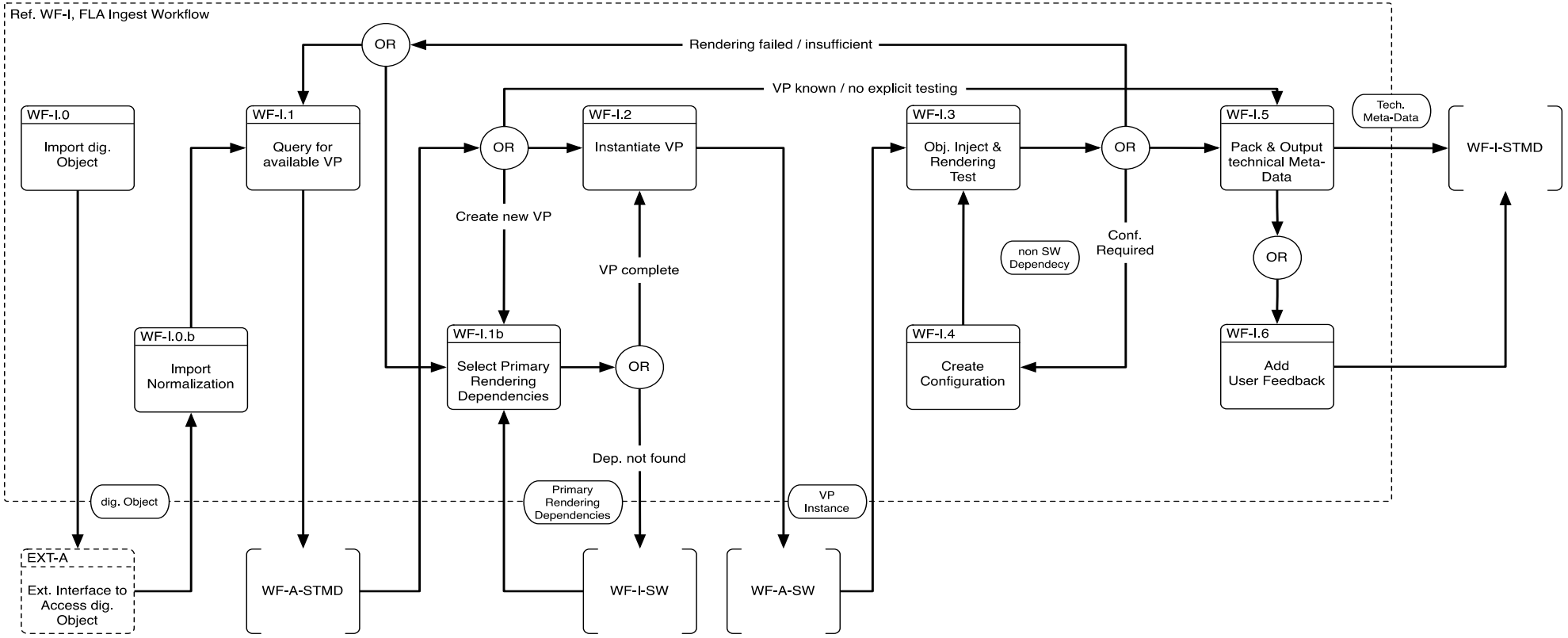
bwFLA – Projektziele

- Fokus des Projekts:
  - Interaktive Documente
    - Z.B. e-Learning – Objekte mit komplexen inneren Strukturen, interaktive DOCs mit Macros
  - DOs mit komplexer innerer Struktur erfordern die Analyse eines komplexen System setups (“rendering”)
    - Z.B. Databank mit einem komplexen view-layer (wie CMS)
    - → authentisches vollständiges Re-enactment solcher DOs in ihrer Originalumgebung
  - System Preservation
    - Sammlung von unabhängigen DOs
    - Z.B. wissenschaftliche Daten mit “Eigenbau”-Analysewerkzeugen und problembezogenem Workflow

bwFLA – Projektziele

- Ziel: Bereitstellung eines Framework und von Workflows, um den langfristigen Zugang zu digitalen Objekten via Emulation zu ermöglichen
- bwFLA – ein nutzerzentrierter Zugang
  - Nutzerwissen benutzen um alle Bedingungen an die Originalumgebung des DO zu erkennen, so dass das “rendering” eine vollständige Umgebung ergibt, welche frei von Abhängigkeits-Konflickten ist.
  - Bewahrung von Wissen über die Installation, Nutzung und Konfiguration der Objekte und Umgebungen
  - Liefern einer “Vorschau” auf die Qualität des Rendering für Objekte und Umgebungen

bwFLA – Ingest Workflow



bwFLA – Ingest Workflow

bwFLA
Functional Long-Term Access

Rendering the Object.

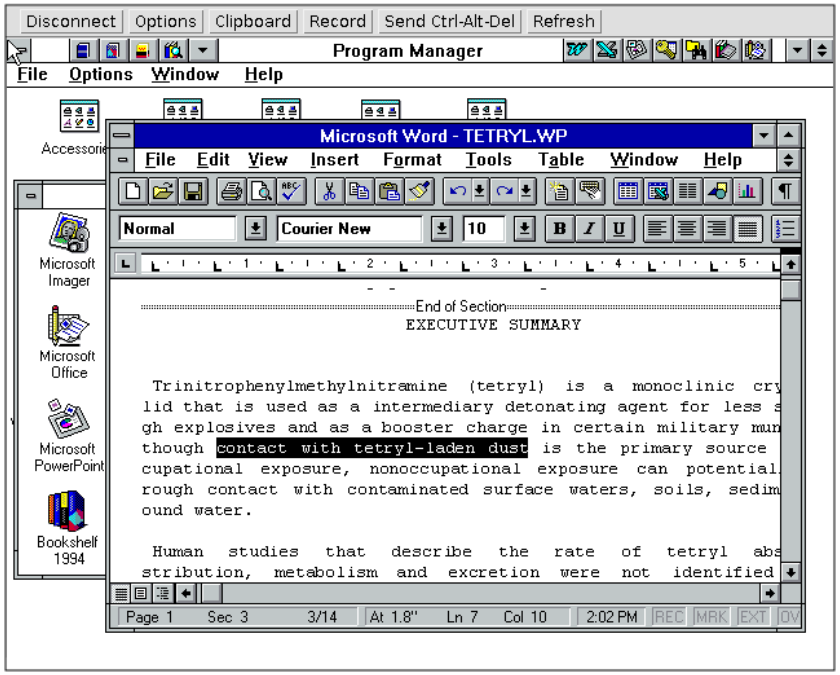
At this step all the dependencies which constitute the rendering environment for your object are resolved. The object will now be injected into the rendering environment on an emulated storage device. Please perform the steps required to render the object by interacting with the emulated rendering environment.

Afterwards, please provide your feedback by confirming or rejecting the rendering operation. Depending on your choice you will be redirected to a corresponding workflow step.

Viewpath:

- Qemu 0.15.1
- MS Windows 3.11
- GhostScript 4.03
- MS Word 6.0
- Object Rendering

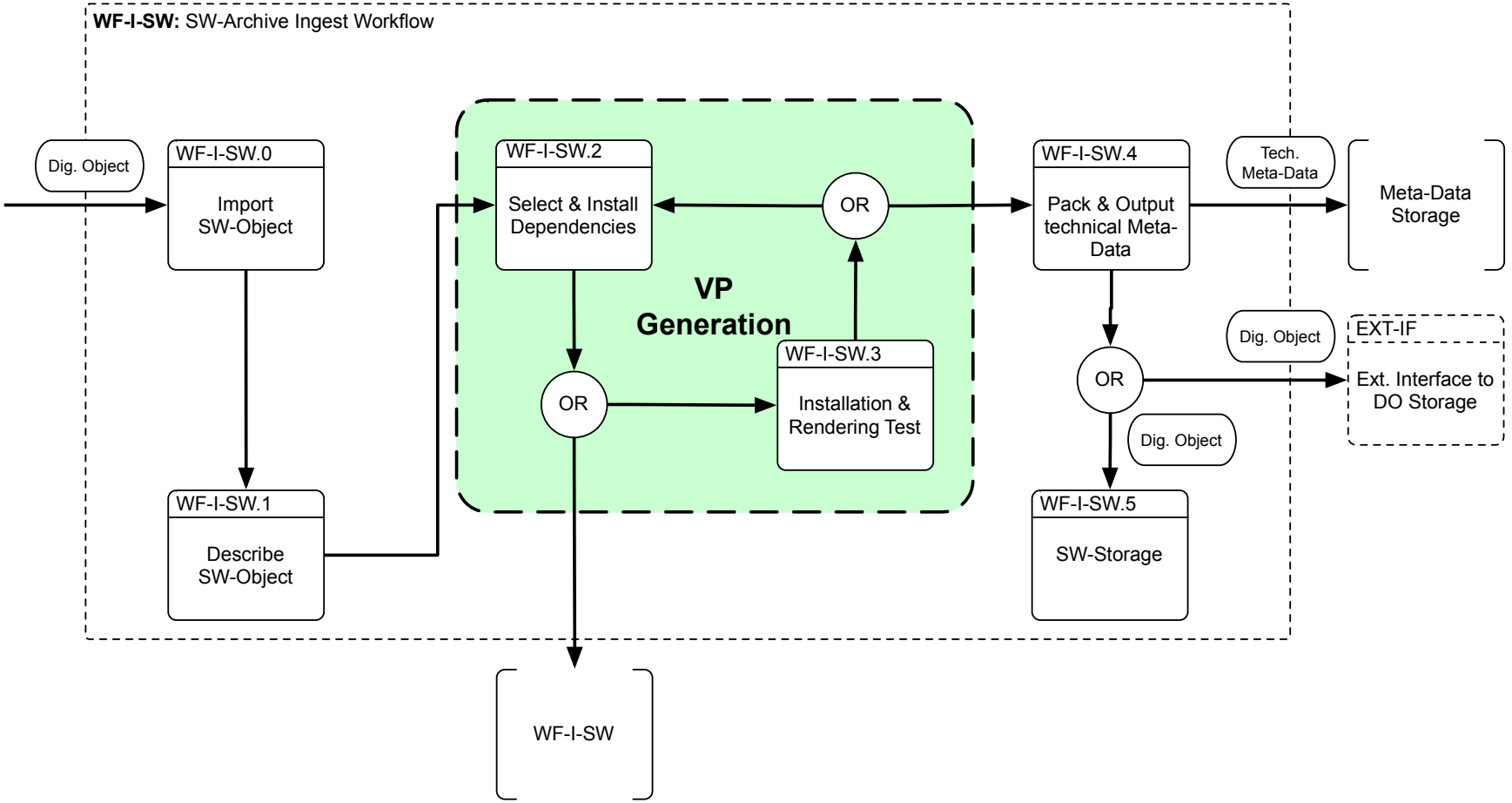
Confirm Rendering
Reject Rendering



Software Archiv

- Viewpath generation workflow
  → bwFLA liefert eine webbasierte Umgebung, die einen strukturierten Installationsprozess auf einer Referenzmaschine erlaubt
  → enthält somit verifizierte und technisch vollständige Metadaten über die Softwarekomponenten
- Emulation als ein Web service
- Automatisierte Viewpath Einrichtung/Generierung
  -Installation der SW Komponenten wird als abstrakte interaktive Workflow-Beschreibung mitprotokolliert und ausgewertet (recorded during ingest)
  → automatisierte VP Erstellung für gewisse Umgebungen möglich
- Bewahrung von ausgewählten Umgebungen
- Automatisiertes Testen von Emulatoren
  -Erlaubt automatische Ausführung von gespeicherten Emulatorumgebungen im Falle von Emulator Upgrades.
  - Um Verfälschungen von vornherein ausschließen zu können

Software Archive Workflow



Software Archive: Installation of SW

bwFLA Functional Long-Term Access

Installing Dependency.

At this step you are able install a chosen dependency object for your software component. An emulated environment will appear and you will be able to install your dependency object injected on a storage device.

After the installation finishes please provide your feedback by confirming or rejecting this dependency. Depending on your choice you will be redirected to the appropriate step.

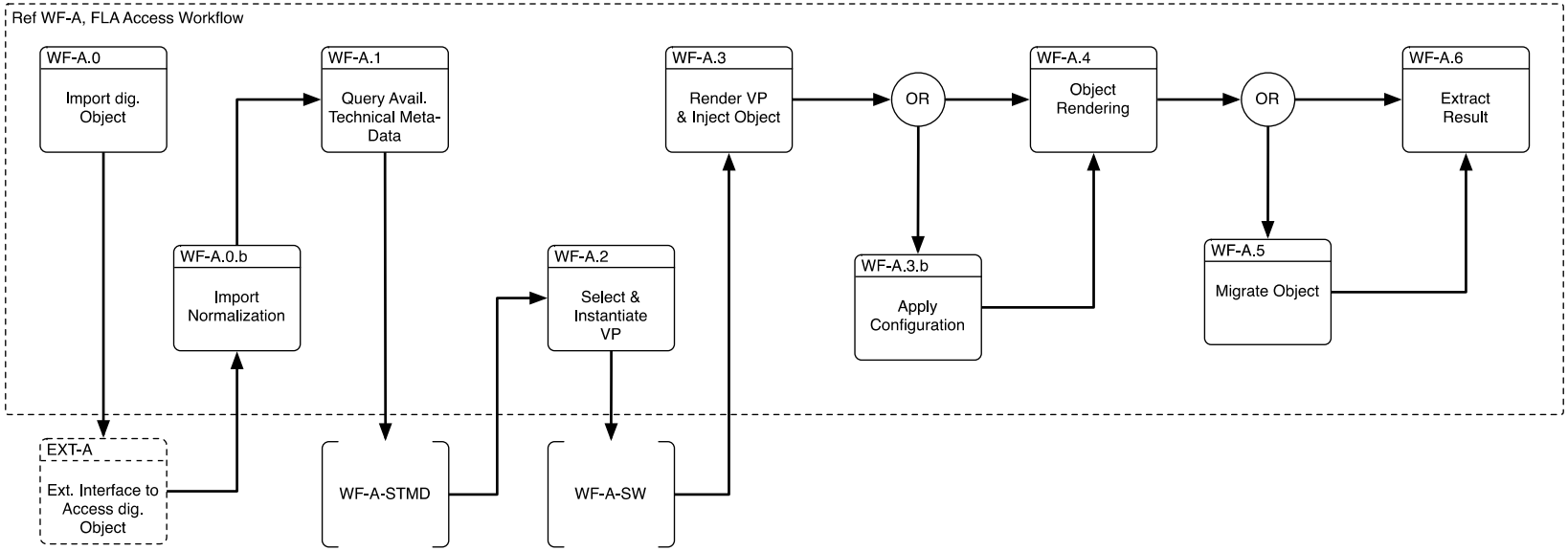
You can view a detailed description of the current viewpath component by pressing on "Object Details" button.

Viewpath:

- Qemu 10.06
- MS Windows 95
- GhostView 5.50
- Comic Sans MS Fonts
- PDF Creator 1.2.3
- MS Office 95

Object Details
Confirm Installation
Reject Installation

bwFLA – Access Workflow



Blick in die nahe Zukunft

- Emulation ist auf dem Weg zu einer reifen und akzeptierten digitalen Archivierungsstrategie
- Emulation ist für gewisse Objekttypen unverzichtbar
- Vorgestellte Workflows erscheinen nur auf den ersten Blick mühsam und teuer
  - De facto wenig Magie notwendig (noch)
  - Workflows benötigen menschliche Interaction

    → preservation Planung auf Emulator fokussiert

    → „authentisches“ Rendering wird garantiert

    → Bewahrung von Nutzerwissen

Folgen ...

- Erkenntnis:
  - Damit es langfristig funktioniert, muss die Langzeitarchivierungs-Community aktiver bei der Entwicklung von Open Source Emulatoren werden
    - Bisher kein Kerngeschäft der Bibliotheken
  - Permanentes und frühzeitiges Testen von alter Hardware und Systemen, die nicht mehr herstellt werden
- Zusammenarbeit und verteilte Ansätze notwendig:

→ share the knowledge

→ share the pain

→ share the cost

→ share the gain ☺

bwFLA
Functional Long-Term Access

University of Freiburg
- Concept, project management and development

University of Ulm
- Sci. data and integration of university libraries

Bibliotheksservice-Zentrum Baden-Württemberg
- Integration, meta-data and coordination of libraries

Landesarchiv Baden-Württemberg
- Archive-integration, meta-data

KÜdKA KIT Karlsruhe / HfG Karlsruhe
- Integration with HfG Karlsruhe archives (digital art)

UNI FREIBURG

Functional Long-Term Archival and Access through Emulation

Leitung: Klaus Rechert, Albert-Ludwigs Universität Freiburg

Zusammenfassung (1)

- Der FLA-Ansatz erweitert die „Datenreaktivierungsmöglichkeiten“ erheblich
  - Löst aber leider nicht alles
- Offene Fragen bleiben:
  - Online-Umgebungen, zeitliche Abhängigkeiten
  - Digitalkunst
    - Oft nur Einzelstücke mit künstlerischen Erweiterungen
      - Wir sind da dran!
  - Rechtsfragen: darf ich eine Hardware emulieren?
    - Wie komme ich an die Spezifikationen?
      - Die übrigens vorhanden sind – schon bei der Entwicklung der Hardware
    - Erhalte ich die notwendige (proprietäre) Software?
      - Offenbar sind die Niederlande schon etwas weiter als wir

Zusammenfassung (2)

- Digitale Langzeitarchivierung benötigt in Zukunft zahlreiche Spezialisten
  - Rechenzentren können vermutlich die Binärdaten (ohne Verständnis des Inhalts) sauber in die Zukunft retten
  - Bibliotheken können vermutlich nur noch einen Teilbereich des „wissenschaftlichen Umfelds“ bedienen.
  - Weitere Spezialisten in unterschiedlichen Einrichtungen werden sich um die Datennutzbarkeit in ihren Teilbereichen kümmern müssen
    - Forschungsinfrastrukturen werden in einem viel größeren Zusammenhang bedeutend
    - Die Kosten müssen in Zukunft von vornherein bedacht werden – sonst ist alles verschwunden
      - INF-Projekte haben derzeit einen schweren Stand!

Zusammenfassung (3)

- Forschungsinfrastrukturen müssen aktiven Einfluss ausüben:
  - Würdige Daten sind nur solche, die ohne Verlust an Information archiviert werden können
    - Keine proprietären Umgebungen erlaubt ?!
  - Wie sehen die Standards aus?
    - Abstraktionsebenen?
  - Was können wir überhaupt „stemmen“?
- Was können wir uns leisten?
  - Finanziell?
  - Technisch?
  - Personell?
- Auf DINI warten neue Mitglieder