Forschungsdatenmanagement – sind Universitäten und ihre Infrastruktur-Einrichtungen darauf vorbereitet?

Gerhard Schneider
direktor@rz.uni-freiburg.de
Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Forschungsdaten

- An Universitäten wird mit Daten recht sorglos umgegangen
  - Nach erfolgreichem Experiment werden nur die zielführenden Daten aufbewahrt
    - Andere müssen in anderem Kontext die Experimente wiederholen – teuer
  - Aufbewahrung nicht unbedingt professionell
    - In Schubladen auf DVDs und Festplatten
  - Auffindbarkeit eher selten gesichert
  - Aus Sicht des Experimentators ist das aber völlig in Ordnung
    - Ziel erreicht (Promotion, Publikation)
- Frage nach der wissenschaftlichen Integrität
  - DFG: 10 Jahre nach Projektende (=22 Jahre bei SFB)
- Erste Antwort der Bibliotheken/RZs: Repositorien

Forschungsdaten und ihr Wert

- Aufbewahrung von Daten über einen langen Zeitraum
- **Lang?**
  - 100000 Jahre? (Atommüll)
  - 5000 Jahre? (Archäologie)
  - 500 Jahre? (Bibliotheken, Archive)
  - 100 Jahre? (Mikrofilmarchive)
  - 5 Jahre? (Digitale Daten im Computer)
- Warum aufbewahren?
  - Daten haben einen Wert
    - Und sei es nur die Arbeitszeit, die zur Gewinnung nötig war
  - Daten sind einzigartig
    - Wie Wetterdaten, Satellitendaten

Forschungsdaten - Typen

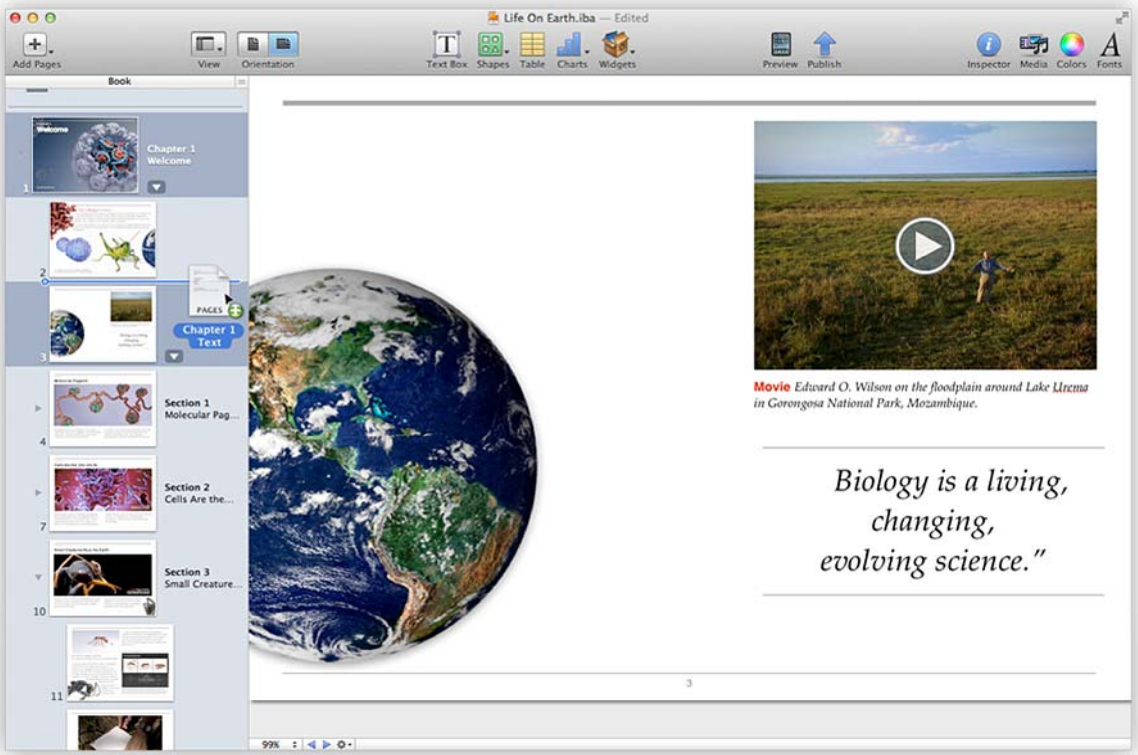
- Welche Daten?
  - *Klassisch*: statische Daten wie ASCII-Texte, Bilder (Scans), usw.
    - Also Daten, die zur Darstellung wenig Technik benötigen
    - Typischerweise in Repositorien / prägen unser Denken
  - *Schwieriger*: statische Daten, die mit proprietärer Software erzeugt werden
    - MS Word, ppt,
    - Zur Darstellung ist proprietäre Software notwendig
      - Ist diese verfügbar?? Lizenzrechte?
  - *Hart*: Daten, die nur mit einem Darstellungsprogramm untersucht werden können
    - Multimedia, Datenbanken, Lernsoftware
  - *Unmöglich* (?): vernetzte, voneinander abhängige Daten

Forschungsdaten à la Apple ….

- *With galleries, video, interactive diagrams, 3D objects, and more, these books bring content to life in ways the printed page never could.*

Ignorieren??

- Möglicherweise ist dies die Zukunft!
- Entweder man liest 10 Seiten „Bleiwüste“, um die Reproduktion der Zellen zu verstehen - oder man schaut sich ein kurzes (und gutes) Filmchen an.



Wie archiviert man das?
Repositorien winken ab!

Die Wissenschaft greift alles auf, was (karrieremäßig) voranbringt

Forschungsdaten aufheben

- Technisch ist die Aufbewahrung der Daten nicht weiter schwierig
  - Das kann ein RZ – bitstream preservation durch Kopieren.
  - Wiederfinden – das kann typischerweise eine UB
- Aber wie interpretiert man die Daten, wenn man sie braucht?
  - Word 2.0 Dateien? WordPerfect Dateien?
- Strategie 1: Migration
  - Immer wieder umwandeln in ein neues Format
  - Flüsterpost-Problem
  - Hoher Aufwand, aber nur auf Verdacht – Nutzung unklar
  - In genau definiertem Umfeld erfolgversprechend

Forschungsdaten-Management

- Wer kümmert sich um die angefallenen Forschungsdaten?
- Relativ neu bei der DFG: INF-Projekte
  - Mittel für Infrastruktur in der Ergänzungsausstattung
  - Wer aber „nimmt“ die Infrastrukturmittel?
  - Kann ein Wissenschaftler Infrastruktur?
    - Insbesondere den zugehörigen Zeithorizont?
- Große Datenmengen – big data
  - Werden oft ignoriert
- Strukturierung der Forschungsdaten?
  - Organisationsfragen

E-Science in Freiburg - RZ und UB als strategische Partner für SFBs und Drittmittelprojekte

SFB 1015
„Muße. Konzepte, Räume, Figuren“

- Gestartet: 1.1.2013
- Teilprojekte: 18*
- Disziplinen: 12
- Stellen für INF: 1 E13, 1 E9 (je 50% RZ/UB)

SFB 1110 (in Begutachtung)
„The Dynamics of Language Spaces“

- Startet: 1.7.2013 (erhofft)
- Teilprojekte: 14*
- Disziplinen: 6
- Stellen für INF: 0.5 E13, 0.5 E9

SlaVaComp
„COMputergestütze Untersuchung von Variabilität im KirchenSLAvischen“

- Gestartet: 1.1.2013
- Gefördert durch: BMBF
- Laufzeit: 3 Jahre
- Stellen für „INF“: 1 E13 (100% RZ)

Tambora

- Gestartet: 1.7.2010
- Gefördert durch: DFG
- Laufzeit: 2 Jahre
- Stellen für INF: 100% UB

Gefördert durch die DFG Deutsche Forschungsgemeinschaft

* = inkl. INF-Projekt

INF-Projekte in SFBs

Schwerpunkte:

- Aufbau der passenden virtuellen Forschungsumgebung
- Plattformen für optimale Kommunikation

(Collaborative tools, Workflow tools bis hin zur automatischen Publikation + OpenAccess)

- Unterstützung der Forschung durch Eigenentwicklungen

(SlaVaComp → TextGrid Plugins und spezifische Tools)

- Datenrepositorien (UB, RZ) mit funktionalen Aspekten zur Langzeitarchivierung (bwFLA)

Herausforderungen

- Integration der INF-Projekte in Servicelandschaft der Universität

(INF-Mitarbeiter „eingebettet“ in die Organisation, aber zu 100% dem SFB bzw. Projekt zugeordnet)

- Einbringen von Betriebswissen der zentralen Einrichtungen in Projekte
- Koordination der Informationsflüsse zwischen INF-Mitarbeitern untereinander und Mitarbeitern der Einrichtungen

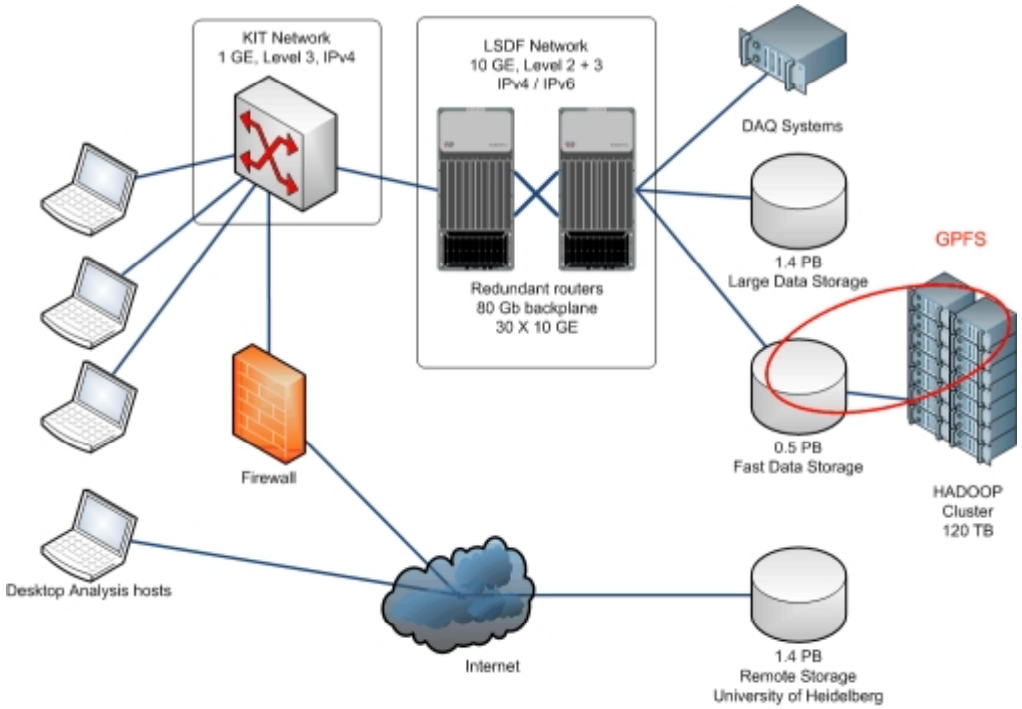
E-Science in Freiburg:

INF-Projekte als Innovationsmotor für zentrale Einrichtungen

Aktivitäten in Ba-Wü

- Notwendigkeit eines funktionsfähigen Frameworks erkannt
- Mit LSDF existiert im Land bereits das „Datengrab“, das die bitstream-preservation garantiert beherrscht

Currently, the **LSDF** hosts around 1 PB of data and provides **hadoop** and cloud resources to over 100 scientists at 17 institutes. The LSDF will be part of the federated IT infrastructure

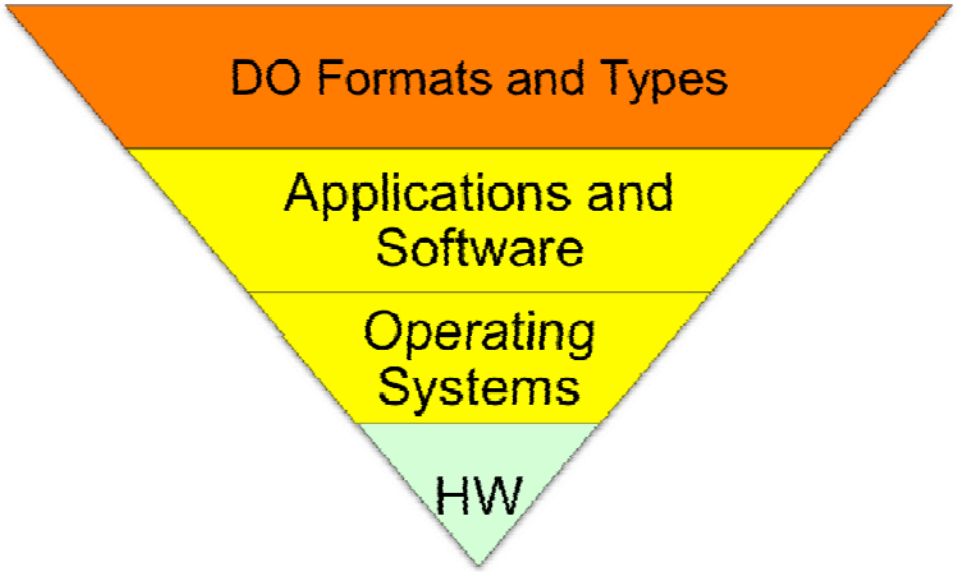


Aktivitäten in Ba-Wü

- Big data: bitstream ist „gelöst“
- Aktuell: Entwicklung von Workflows
  - zur Bereitstellung archivierbarer Laufzeitumgebungen
  - zur Aufnahme von Forschungsdaten
  - Und deren Auslieferung ganz oder in Teilen
  - Ergänzend/komplementär zu bereits existierenden disziplinspezifischen Lösungen
    - Wetterdaten sind nicht unbedingt LSDF-tauglich
- bwFLA als ein zentrales und flexibles Framework (Toolbox zur Reanimation von Daten) einer disziplinspezifischen Lösung

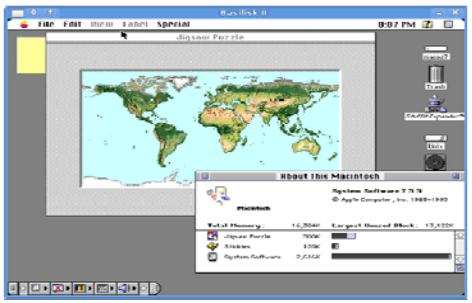
Objektpyramide / Rettung von Forschungsdaten

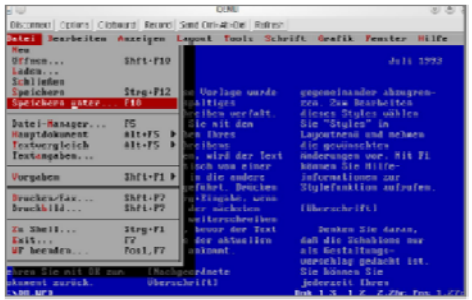
- Unübersehbar viele digitale Objekte in vielfältigen Formaten
- Deutlich weniger Betriebssysteme und nochmal weniger Computer Plattformen

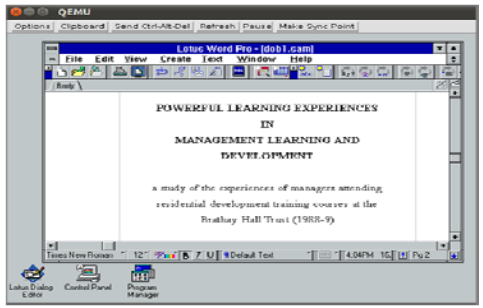


Emulation von Originalumgebungen

- Alternative zur Migration des Objekts – Nutzen seiner Originalumgebung
  - Annahme: Formate funktionieren am besten in ihrer Originalapplikation
  - Software funktioniert in Umgebung am besten, für die sie geschrieben wurde
- Deshalb: Wiederherstellung, Bewahrung von Originalumgebungen







Funktionale Langzeitarchivierung

UNI FREIBURG

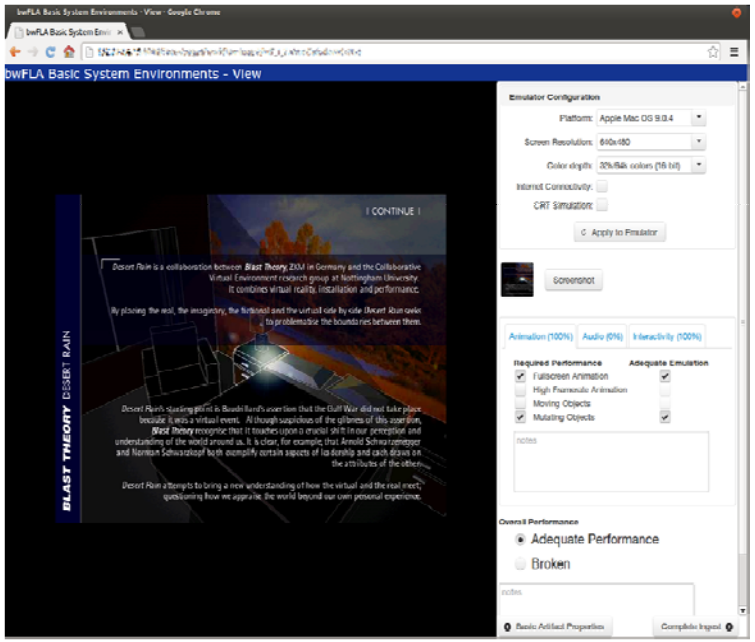
- Forschungen und Entwicklungen der FLA Gruppe in Freiburg für die technische Ebene des Objektzugriffs
- Sehr weit gefasster Objektbegriff
  - Einzelne Dateien verschiedener Formate
  - Gruppen von Dateien
  - Dynamische Objekte wie Datenbanken, Digitale Kunst, Multimedia (im weitesten Sinne)
  - Komplette Systemumgebungen verschiedener Computerarchitekturen

bwFLA Projekt

- Ziel: Schaffung einer skalierenden funktionalen Komponente für Object-Ingest und Access
  - Oktober 2011 – Dezember 2013
  - Service-Infrastruktur ohne eigenen Storage
- Partner
  - KIZ Universität Ulm
  - Rechenzentrum Universität Freiburg
  - HFG Karlsruhe
  - Landesarchiv Stuttgart
  - BSZ Konstanz
  - Neu: UB Freiburg

bwFLA - Use Case 1: Kuratierung Digitaler Kunst

Digitale Kunst als Prototyp nicht kontrollierbarer Forschungsdaten



- Erfassung und Beschreibung der *Transmediale* Sammlung (272 CDs, Zeitraum Mitte 1990er bis Mitte 2000er)
  - Vollständige Umsetzung des Ingest-Workflow
  - Access in verschiedenen Originalumgebungen
  - Inhaltliche Erfassung und Auswertung aller Objekte

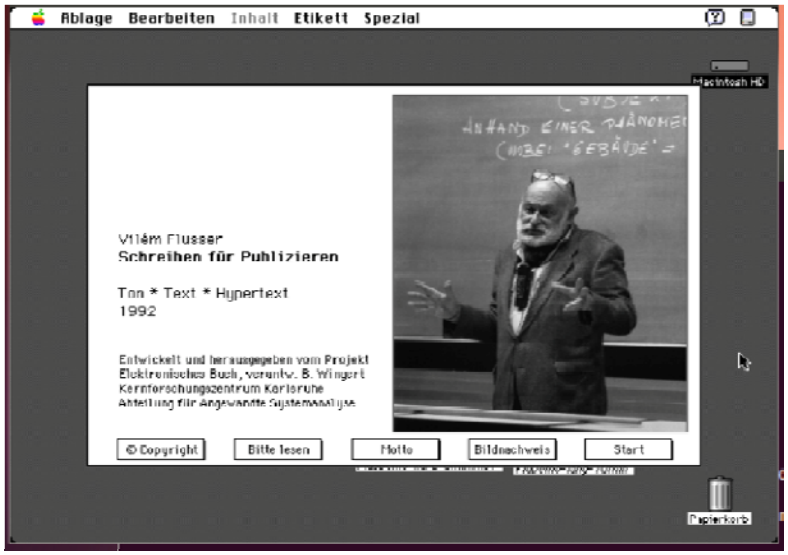
bwFLA - Use Case 2: Full System Preservation

UNI FREIBURG

- Maschinen wichtiger Personen, komplette Forschungs- oder Entwicklungsumgebungen
  - Objekt ist aus sich heraus interessant
  - Keine ausreichende Dokumentation vorhanden
  - Erhalt von Zusammenhängen, Kontext
- Mehrere Beispiele
  - Access Datenbank in Windows 2000
  - Apple Performa von Vilem Flusser
  - OS/2 DB2 Netzwerk von Forschungsprojekt

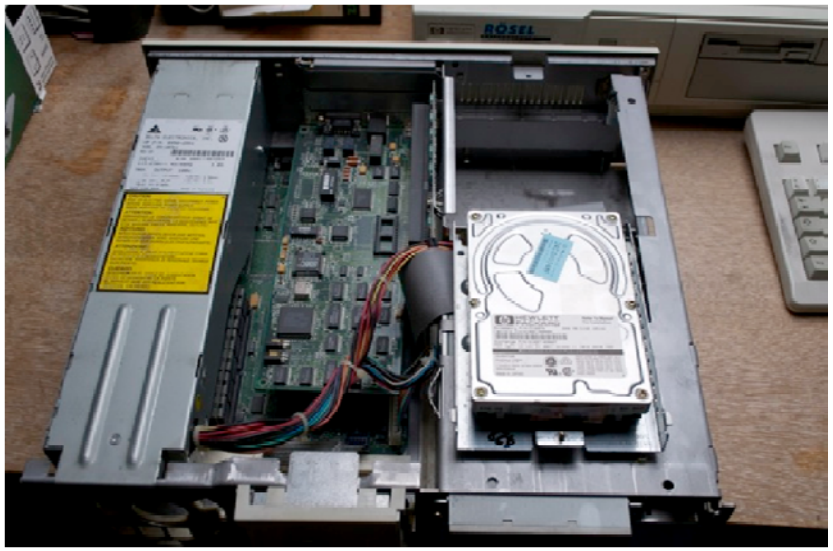
Full System Preservation

- Flusser-Archiv in Berlin besitzt Original-maschine des Philosophen
  - Früher Einsatz von Multimedia, verknüpften Objekten, E-Learning
  - Hypercard-Projekt für Ausstellung



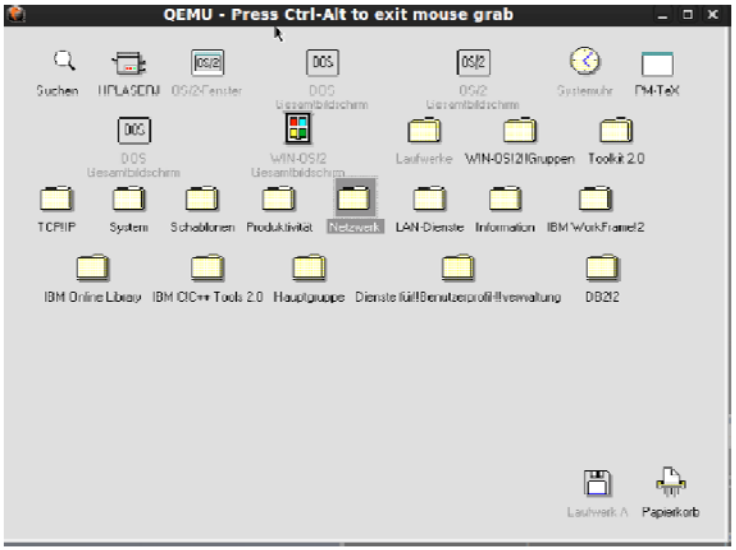
Full System Preservation

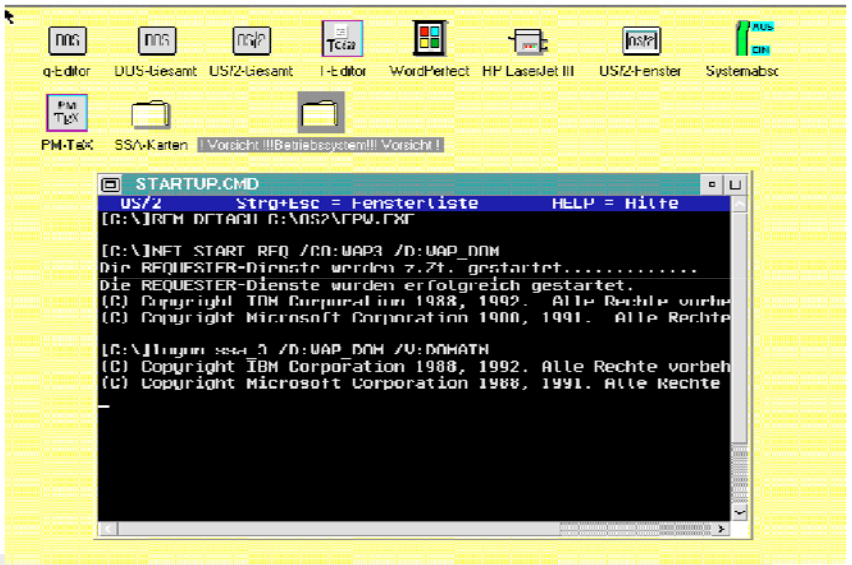
- Bewahrung komplexer Forschungsumgebung
  - Sprachatlas lokaler Dialekte in Südwest-Deutschland
  - Lang laufendes, DFG-gefördertes Projekt mit vielen Nutzern



Full System Preservation

- Client-Server-Architektur
  - Funktionsfähige X86-Systeme aus dem Jahr 1993
  - Festplatten wurden gespiegelt und im Emulator wieder angefahren
  - Tatsächliche Wiedererweckung der Daten





Komplementär: bw-Lehrpool

- Problem: Archivierung der Arbeitsumgebung
  - Idealerweise nicht nötig, aber…
  - was, wenn proprietäre Umgebungen zwingend sind?
  - Elektronische Laborbücher?
  - Spezielle Mikroskop-Treibersoftware?
  - Spezielle proprietäre Auswertesoftware?
- Bw-Lehrpool: Bereitstellung von schnell auszuliefernden, einfach zu konfigurierenden Arbeitsumgebungen
  - Virtuell
  - Einfach zu archivieren
- Standardisierung der Arbeitsplatz-Hardware
  - Per Hardware: bw-PC
  - Per Software: Ablösen von der realen Umgebung

Herausforderungen

- Einführung der Frameworks in das akademische Leben
  - Geräteauswahl mit Nebenbedingung der Datenarchivierung
  - Reduktion der angeblichen (!) Freiheit zu Gunsten standardisierbarer Forschungsumgebungen
    - Sonst wird es nachher teurer
  - Rücksicht auf Archivierbarkeit als Teil des akademischen Selbstverständnisses
    - Und der akademischen Abläufe
    - Wie: elektronische Abgabe von Doktorarbeiten mit Daten
    - Fragen der Urheberrechte / des Zugriffsschutzes
- Die Arbeit hat gerade erst begonnen
  - Riesige Strukturaufgabe für die Universitäten

Zusammenfassung

- In Ba-Wü existiert ein erstes Konzept, um
  - Daten
  - Arbeitsumgebungen
- so archivieren zu können, dass sie
  - leicht wieder reaktiviert werden können
- Damit ist ein durchgängiger Workflow zur Bewahrung von Forschungsdaten möglich
  - Nächster Schritt: Implementation als möglicher Standard für Doktoranden