



Digitale Sprach-Archive und Langzeitarchivierung

(am)
MPI für Psycholinguistik

Peter Wittenburg

(Humanities orientiert)

1

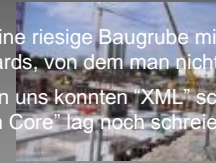


erster Versuch

- vor einigen Jahren einen Vortrag über das sich entwickelnde Web
- mir war gesagt worden, ich solle mal einen Ausblick wagen
- wir waren damals in den Startschuhen – mein Anfangsbild sah daher etwa derart aus – noch handgemalt!

Das Web als eine riesige Baugrube mit vielen vagen Ideen und neuen Standards, von dem man nicht recht wußte, was wird.

Einige von uns konnten "XML" schon buchstabieren.
"Dublin Core" lag noch schreiend in der Wiege.



hatte den Eindruck, daß ich ziemlich an Ihren Interessen vorbeigeredet habe – das Schweigen war zumindest ehrlich

2



zweiter Versuch

- jetzt soll ich Sie 1 Stunde lang informieren/unterhalten/...
- man hofft natürlich immer auf anregende spontane Debatten
- immerhin stehen die MPDL und eSciDoc im Raum
- wir sind inzwischen etwas weiter als damals:
 - wir bauen komplexe Architekturen auf XML als Fundament
 - wir wissen, daß Dublin Core nicht alles ist
- einige erwarten ein Auflösen traditioneller Bereiche
 - das elektronische Publizieren (Mainstream)
 - das Umgehen mit wissenschaftlich relevanten Daten (speziell)
 - sicher: es wird qualifizierte Verweise geben
 - aber: die Prozesse und Inhalte + Formate sind unterschiedlich
 - **Veränderungen werden nur langsam kommen**

3



Übersicht

- **Digitale Archive**
 - **Basis digitaler Archive**
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - Was aufbewahren?
 - Technologische Lösungen
 - Services in der MPG

4



Basis Digitaler Archive I

GSH Treffen
November 2005

Archive

- traditionell speichern Archive physikalische Original-Objekte so, daß sie lange überleben
- das beinhaltet, daß ein Zugriff nur sehr beschränkt gegeben wird
- im Bereich digitaler Daten ist der physikalische Träger nicht mehr wichtig
- warum: Kopien sind nicht mehr verlustbehaftet (Vorsicht!)
- für klassische Archivare eine ungeheure Revolution (IASA – International Association of Sound & Audiovisual Archives)

5



Basis Digitaler Archive II

GSH Treffen
November 2005

Digital Libraries (Community Discussions)

- im Vordergrund steht der Zugriff – die Exploitation der Inhalte
- Archive haben so etwas "staubiges" an sich – sehr unpopulär
- lange Zeit keine Gedanken über LZ-Verfügbarkeit
- LZA kostet Geld und bringt Firmen nichts

Moderne Digitale Archive

- ein Spagat zwischen LZ-Archivierung und direkter Verfügbarkeit
- ein Spagat zwischen Original-Daten und Anreicherungen
- kurz: ein lebendiges dynamisches Repositorium mit LZ-Auftrag

6



Übersicht

GSH Treffen
November 2005

- **Digitale Archive**
 - Basis digitaler Archive
 - **Bausteine digitaler Archive**
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - Was aufbewahren?
 - Technologische Lösungen
 - Services in der MPG

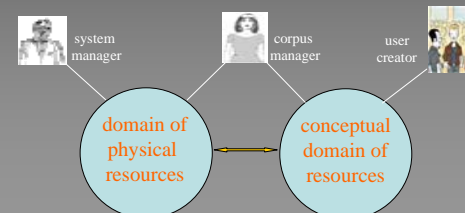
7



Bausteine Digitaler Archive I

GSH Treffen
November 2005

- strikte Trennung physikalischer und logischer Zugangsebene
 - für Bibliotheken alter Hut: Welt der Karten und Welt der Bücher
 - physikalische Welt ist die Welt der System Manager und Archive Manager
 - gerade wegen LZ-Aspekten müssen sie alle Freiheiten haben
 - physikalische Welt (Server, Platten, ...) verändert sich regelmäßig
 - logische Welt (die fachspezifische Beschreibung und Organisation) bleibt
 - Metadaten halten beides zusammen – müssen gepflegt werden
 - Benutzer sieht nur Metadaten mit seiner Terminologie



8

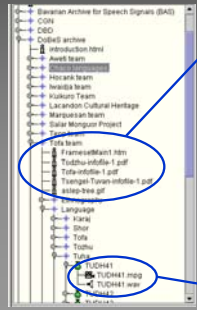


Bausteine Digitaler Archive II

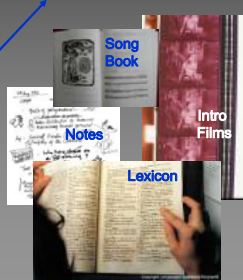
GSH Treffen
November 2005

- Relationsmechanismus
 - vielleicht sogar mal was Neues für Bibliothekare?
 - Relationen vielfältiger Art auf Ressourcen Ebene

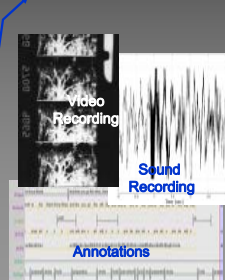
Archiv Organisation



Ebene der Sprache



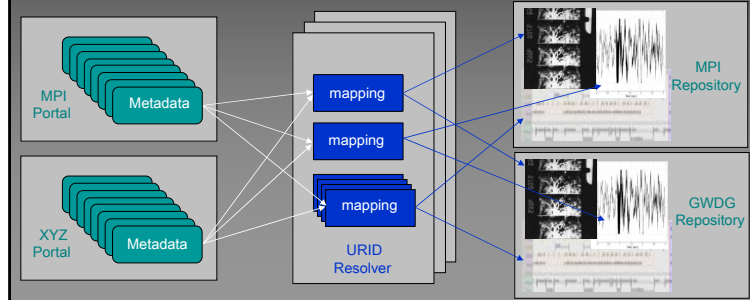
Ebene der Session



Bausteine Digitaler Archive III

GSH Treffen
November 2005

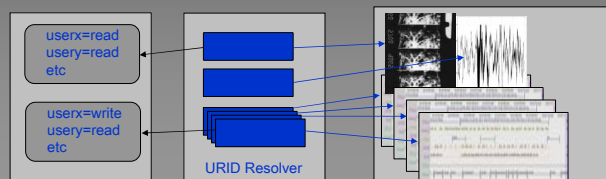
- brauchen Trennung zwischen Objekt und Instanz
 - für Bibliotheken alter Hut: ISBN als Inkarnation vs. alle Kopien
 - brauchen Unique Resource IDs
 - und einen allgemein zugänglichen "Resolving" Mechanismus
 - allgemeinen Präfix und Autorität für Nummern bei zB. Instituten



Bausteine Digitaler Archive IV

GSH Treffen
November 2005

- brauchen Versionierung, Authorisations Information & Authentifizierung
 - außer Versionierung für Bibliothekare nichts Neues (oder?)
 - nichts darf vernichtet werden, aber Annotationen werden geändert!
 - Wissenschaft ist dynamisch
 - URID Information ist zentraler Link zu Authorisationsrekord
 - Authorisation ist nicht an Instanz gekoppelt (anders bei Büchern)
 - Authentifizierung entsprechend normalen Regeln (verschlüsselt, ...)
 - IPR ist sehr wichtig, da wir vertrauliche Daten haben



11



Übersicht

GSH Treffen
November 2005

- Digitale Archive
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- Langzeit-Archivierung
 - Wichtigkeit
 - Was aufbewahren?
 - Technologische Lösungen
 - Services in der MPG

12



MPI Archiv – Stand

GSH Treffen
November 2005

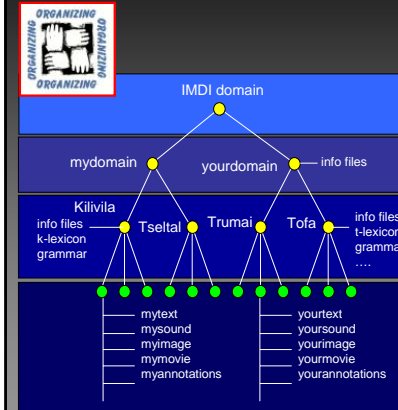
- über 150.000 Objekte (ohne Photos)
- über 11 TB
- pro Jahr etwa 1.5 TB hinzu = 1500 h Audio/Video Aufnahmen
- vollkommen strukturiert mittels gelinkter XML Files basierend auf offenem IMDI Schema
- Verwendung offener Formate – keine Verkapselung dh.
 - bei Speicherung Fokus auf Repräsentation – nicht auf Performanz und Präsentation
 - alles in archivierbarer Form
 - jeder kann seine eigenen Auswertungsprogramme schreiben
 - für schnelle Zugriffe etc interne Verwendung von Datenbanken
- im Sinne des gesagten nahezu komplettes Basis-System im Einsatz
- an URIDs & Versionierung wird momentan gearbeitet

13



MPI Archiv – logische Ebene

GSH Treffen
November 2005



- web-basiertes Katalog-System ist der Schlüssel
- IMDI Metadaten Satz eine Schöpfung von Disziplin Experten
- ist ein quasi Standard (~55 Institute)
- Vokabular wird von ISO übernommen
- stabil seit mehreren Jahren und vollständiger Tool-Satz
- alles konsequent auf offener XML Welt
- ist eine Welt verteilter XML Files – man braucht nur ein URL
- jeder kann also sein eigenes Portal aufbauen



14



Übersicht

GSH Treffen
November 2005

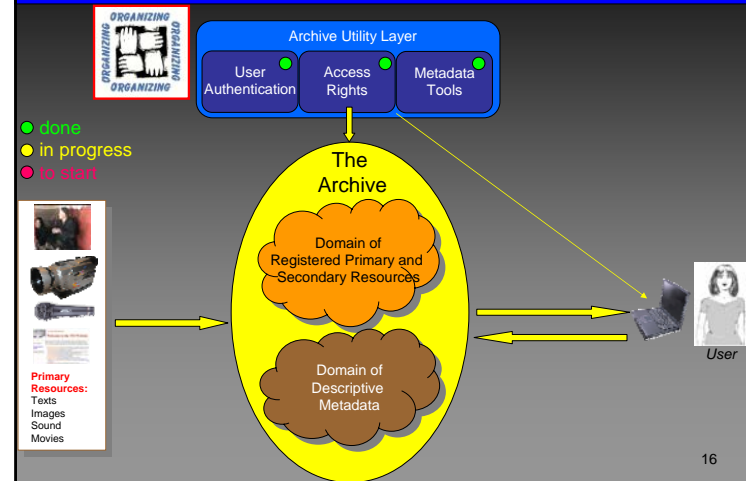
- **Digitale Archive**
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - **Zugang/Import/Verwendung/Export**
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - Was aufbewahren?
 - Technologische Lösungen
 - Services in der MPG

15



MPI Archiv – Tools

GSH Treffen
November 2005



16



MPI Archiv – Metadaten Tools

GSH Treffen
November 2005

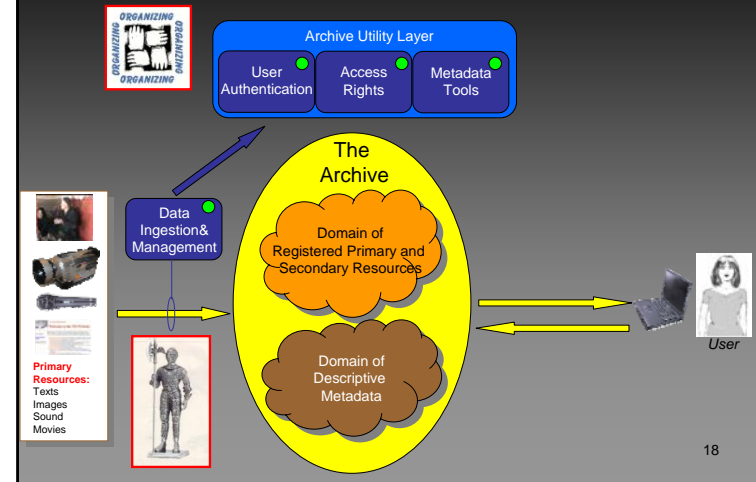
- professioneller Editor zur Erzeugung korrekter IMDI XML Files
- “native” IMDI XML Browser
- Support für HTML Browsing – “on the fly” Transformation
- geographisches Browsen via Google Earth (in der Mache)
- strukturierte Suche in XML and HTML Domänen
- unstrukturierte Suche in XML and HTML Domänen
- unstrukturierte Suche via Google
- Support OAI PMH Gateway für Dublin Core Service Provider
- **Access Management System (Prozeduren, Delegation, effizient)**
- **Trennung interner und externer Benutzer**

17



MPI Archiv – Tools

GSH Treffen
November 2005



18



MPI Archiv – der Torwächter

GSH Treffen
November 2005

LAMUS (Language Archive Management and Upload System)

- ein Content Management System mit Kenntnissen über LR
- **Überwindung des Archive Manager Bottlenecks**
- vollständig web-basiertes Interface
- daher jetzt auch Öffnung des MPI Archives für Externe
- Definition von Korpusstrukturen
- Upload von Metadaten und Ressourcen
- Linken von allem
- Workspace; Format und Konsistenz Checks
- Index Erzeugung (für alle Formen des Zugriffs relevant)
- **Erzeugung von URIDs und Versionierung in Arbeit**



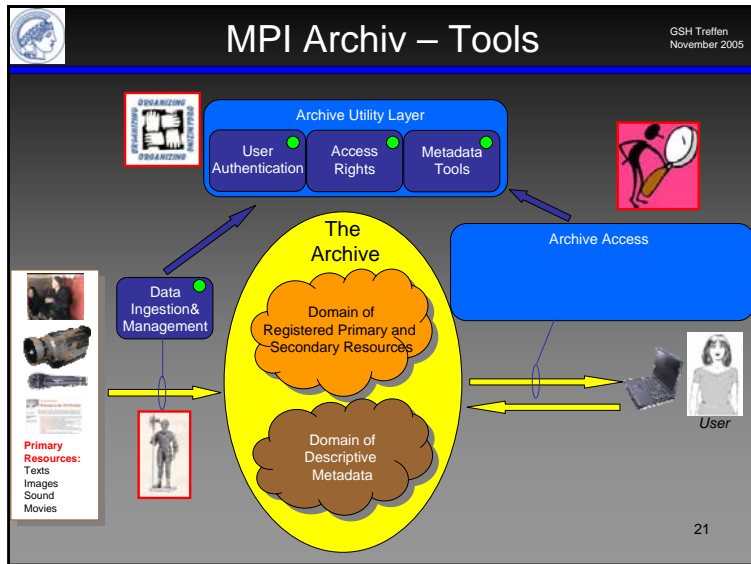
19



MPI Archiv – der Torwächter

GSH Treffen
November 2005

20



MPI Archiv – Zugriff I

GSH Treffen
November 2005

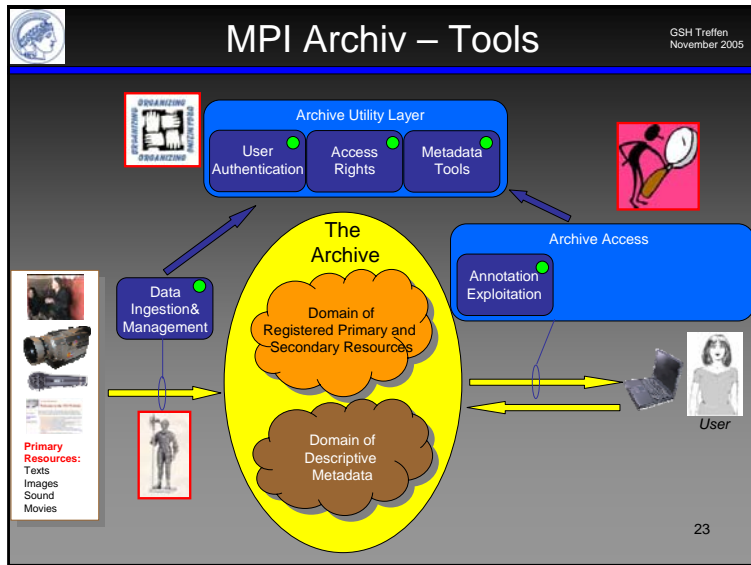
web-basierte Einfachst-Methoden

- Browse oder Suche nach einer Resource
- direktes Herunterladen (keine Kunst)
- direkte Visualisierung (Video erfordert Streaming Server/Client)
- **Kopieren ganzer Sub-Archive inklusive der Organisation** (sehr wichtig für Corpora von Sprach-Communities)

web-browser & plug-ins

play

22



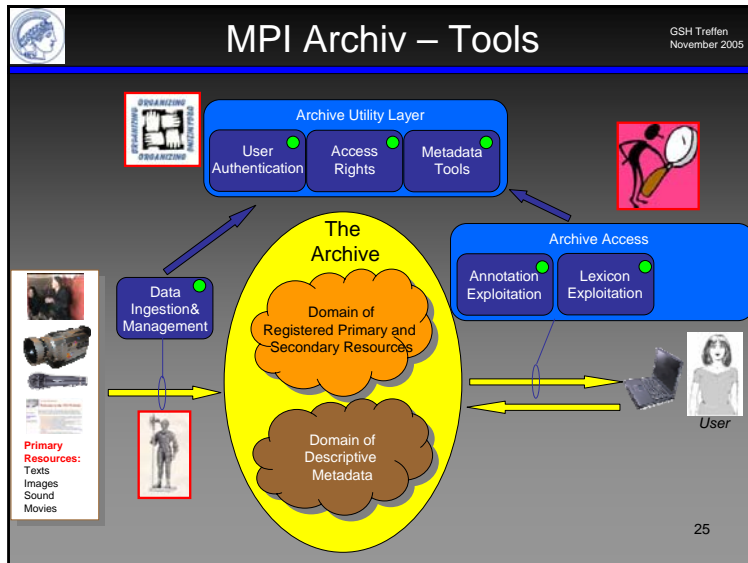
MPI Archiv – Zugriff II

GSH Treffen
November 2005

Zugriff auf komplexe miteinander verbundene Objekte

- **Schritt 1: Benutzer wählt ein annotiertes Medien-Objekt** besteht aus mehreren Video/Audio/Transcription Tracks

24



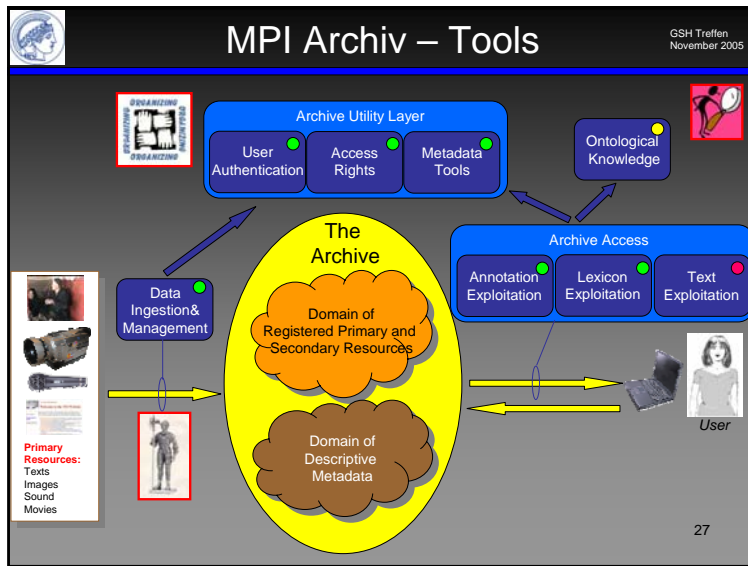
MPI Archiv – Zugriff III

GSH Treffen
November 2005

Zugriff auf komplexe miteinander verbundene Objekte
kollaboratives Arbeiten und (semi) automatisches Mergen

- Schritt 1: Benutzer wählt ein multimediales Lexikon besteht aus Lexikon und mehreren Video/Audio Clips/Refs

26



MPI Archiv – Zugriff IV

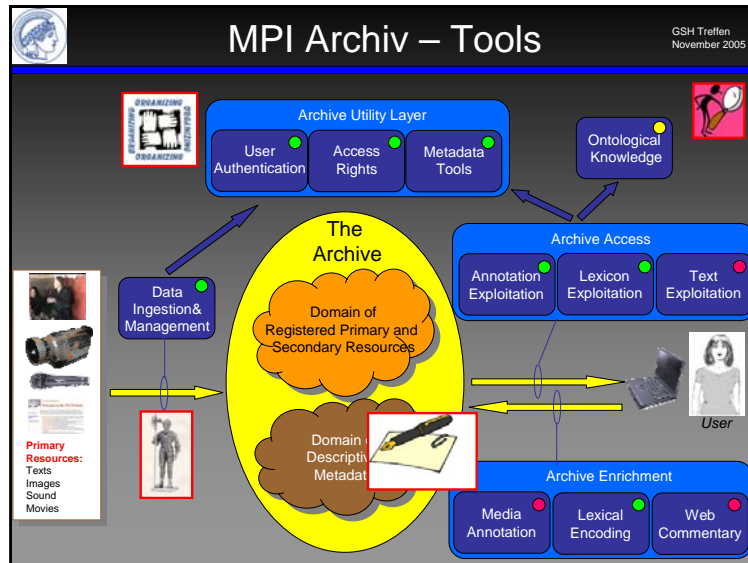
GSH Treffen
November 2005

Idee: Benutzer erstellt sich seine virtuelle Arbeits-Domäne aus verschiedenen Sub-Corpora und sogar Archiven

Grundproblem

- verschiedene Terminologien
- bedarf Ontologien
- mapping: **verb = werkwoord**
- arbeiten an Lösungen
- top-down (ISO) und bottom-up Ansätze

28



MPI Archiv – Tools

GSH Treffen
November 2005

- arbeiten an einem kompletten Archiv Support
- das meiste an Basis-Funktionalität ist vorhanden
- trotzdem noch viel zu tun

30

Übersicht

GSH Treffen
November 2005

- **Digitale Archive**
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - Was aufbewahren?
 - Technologische Lösungen
 - Services in der MPG

31

Herausforderungen

GSH Treffen
November 2005

- flexible Commentary, Kollaborations und Relations Umgebung
 - Verhindern des Junk Problems
 - keinesfalls den vorhandenen Inhalt verändern
- Archiv Federationen – transparente virtuelle Domäne
 - keine Institutsgrenzen etc
 - gemeinsame logische Ebene
 - gemeinsame URID Domän
 - single Sign-On, single Identity
 - distributierte Authentifizierung & Autorisierung
 - sind Mitglied in DELAMAN (Digital Endangered Languages and Music Archive Network)
 - sind Koordinator im DAM-LR Projekt (Distributed Access Management for LR)
- Lösung des Interoperabilitäts Problems
 - Character Encodierung
 - Struktur und Formate
 - Semantik der Begriffe

32



Übersicht

GSH Treffen
November 2005

- **Digitale Archive**
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - **Wichtigkeit**
 - Was aufbewahren?
 - Technologische Lösungen
 - Services in der MPG

33



Langzeit-Archivierung: warum

GSH Treffen
November 2005

- verschiedene Gruppen und Anforderungen
 - Vorhaltung wissenschaftlicher Nachweise (10 Jahre)
 - Publikationen wegen Referenzen etc (?? Jahre)
 - Speicherung nicht-wiederbringbarer Aufnahmen (unendlich)
- wir müssen Diversität für unsere Nachkommen dokumentieren
 - vieles wird in rasantem Tempo verschwinden
 - zB. Diversität der Kulturen und Sprachen
- D. Schüller (UNESCO):
 - 80% unserer Aufnahmen über Kulturen und Sprachen sind akut bedroht
- Publisher
 - kein Interesse an Langzeit-Aufbewahrung
 - kostet viel Geld, kein Business Modell vorhanden



Übersicht

GSH Treffen
November 2005

- **Digitale Archive**
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - **Was aufbewahren?**
 - Technologische Lösungen
 - Services in der MPG

35



Langzeit-Archivierung: was

GSH Treffen
November 2005

- spreche nur über digitale Daten
- Bewahrung des Inhaltes – nicht des Trägers
 - Kopieren geht ohne Verlust (Vorsicht bei komprimierten Daten)
- Bit-Streams (reine digitale Basis Information)
 - Fokus aller Anstrengungen
- interpretative Informationen (technische Encodierung)
- Bedeutung der Codierungen (wissenschaftliche Encodierung)
- organisatorische Informationen (Relationen)
- Sicherung der Algorithmen / allgemeine Emulatoren (vielleicht zeitweilige Lösung für PDF etc)
- sprechen wir über 100+ Jahre?
ein Heer von Daten-Archäologen wird sich über unsere Daten hermachen, wenn sie denn als Bit-Streams überleben
- Konsistenz wird Bereitschaft zum Migrieren erhöhen

011001010100001010110100101010

創新精神





Übersicht

GSH Treffen
November 2005

- **Digitale Archive**
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - Was aufbewahren?
 - **Technologische Lösungen**
 - Services in der MPG

37



LZA Technologie

GSH Treffen
November 2005



- **CDROM/DVDROMs etc sind keine Lösung!**
- Methoden sind ziemlich dürftig und zT. uralt
 - Zentralisierung
 - Austausch der Daten was das Zeug hält
 - ständige Migration zu neuen Technologien (alle Ebenen)
 - **niedrige Kosten sind entscheidend!**
- das MPI kann sich gegenwärtig auf 7 Kopien verlassen
MPI (2), GWDG (2), RZG (2), Evol. Anthropologie (1)



38



Übersicht

GSH Treffen
November 2005

- **Digitale Archive**
 - Basis digitaler Archive
 - Bausteine digitaler Archive
 - das digitale Archiv am MPI
 - Zugang/Import/Verwendung/Export
 - gegenwärtige Herausforderungen
- **Langzeit-Archivierung**
 - Wichtigkeit
 - Was aufbewahren?
 - Technologische Lösungen
 - **Services in der MPG**

39



LZA Angebot in der MPG

GSH Treffen
November 2005

- haben hervorragende Rechenzentren
- GWDG und RZG bieten Services für alle MPLe
- Unterhalten jeweils 2 Kopien dh. Redundanz
- MPG gibt institutionelle Garantie für 50 Jahre !
gilt natürlich nur zur Absicherung des Angebotes der RZ
- BAR "kontrolliert" die Politik der RZ
- gegenwärtig keine expliziten Kosten für MPLe
- Lackner (IPP):
jetzige Datenmengen machen nach einem Technologiezyklus (10 Jahre) allenfalls 10% der dann entstehenden Daten aus

40



vielen Dank für Ihre Aufmerksamkeit