



Überblick: Arbeitspaket und Architektur

Begutachtung des WissGrid AP 3
28. Januar 2010, AIP Potsdam

Jens Ludwig
SUB Göttingen



Bundesministerium
für Bildung
und Forschung



- Auftrag
- Vorgehen
- LZA-Verständnis
- Architektur
- Grid-Community-Unterstützung
- Weiterer Ablauf



- Auftrag
- Vorgehen
- LZA-Verständnis
- Architektur
- Grid-Community-Unterstützung
- Weiterer Ablauf



Auftrag des Arbeitspakets

- formales Ziel laut Antrag:
 - Generische LZA-Architektur für Forschungsdaten in D-Grid
 - Umsetzung der Architektur
 - Modulare LZA-Dienste
 - Grid-Forschungsdatenarchiv
 - LZA-Leitfäden/-Blaupausen
- inhaltliches Ziel:
 - Grid-Communities starke und attraktive Instrumente für ein langfristiges und nachhaltiges Datenmanagement anbieten!
 - Dadurch ein weiteres Argument für D-Grid schaffen...



Stand im AP-Projektplan

M06: Konzept für eine generische Grid-LZA Architektur

M08: 1. Iteration Spezifikation: LZA-Dienste und Repository

M09: Begutachtung Architekturkonzept

(Deliverables zu M06 und M08 Teil der Begutachtung)

M12: 2. Iteration Spezifikation: LZA-Dienste und Repository

M24: 1. Iteration Umsetzung LZA-Dienste und Repository

...

M26: Workshop zu LZA-Leitfäden

M36: 2. Iteration Umsetzung LZA-Dienste und Repository

...



- Auftrag
- **Vorgehen**
- LZA-Verständnis
- Architektur
- Grid-Community-Unterstützung
- Weiterer Ablauf



Wichtige Erfolgsfaktoren

- Akzeptanz und Integration durch die Grid-Communities
 - Ihre Anforderungen verstehen und aufnehmen, unser bestehendes Wissen weitergeben
 - Generische Entwicklungen, die an Community-spezifische Anforderungen anpassbar sind
 - Wenn möglich Anpassung und Integration begleiten und unterstützen
- Stand der Technik aufnehmen und fortschreiben
 - Viele internationale Vorarbeiten existieren, keine völligen Neuentwicklungen, sondern Grid-Anpassungen
 - Grid mit LZA und LZA mit Grid ist ein z.T. neues Thema



Was wir bisher geleistet haben

- Akzeptanz und Integration durch die Grid-Communities
 - Begleitende AG mit Vertretern der Fachdisziplinen gegründet und erstes Treffen durchgeführt
 - Zwei weitere, nicht im Antrag genannte, aber interessierte Disziplinen: Sozialwissenschaften und Altertumswissenschaften
 - Treffen mit Photonenphysik und Sozialwissenschaften, weitere Anfang Februar
 - Fallstudien im Architekturdokument (werden fortgeschrieben)
 - (Gespräche mit SUN als interessierten kommerziellem Partner)



Was wir bisher geleistet haben

- Stand der Technik aufnehmen und fortschreiben
 - intern: gemeinsames Verständnis und Terminologie
 - 3 Deliverables, die den Stand der Technik analysieren und aufnehmen
 - Konzept für eine generische Grid-LZA Architektur
 - 1. Iteration Spezifikation: LZA-Dienste
 - 1. Iteration Spezifikation: Repository
 - Koorganisation der OGF Digital Repositories Research Group (letzte Workshop im Dezember in London, nächste bei der OGF 28 in München)
 - Paper beim International Symposium on Grid Computing im März 2010, Taiwan
 - Präsentation beim DGI-Metadaten-Workshop

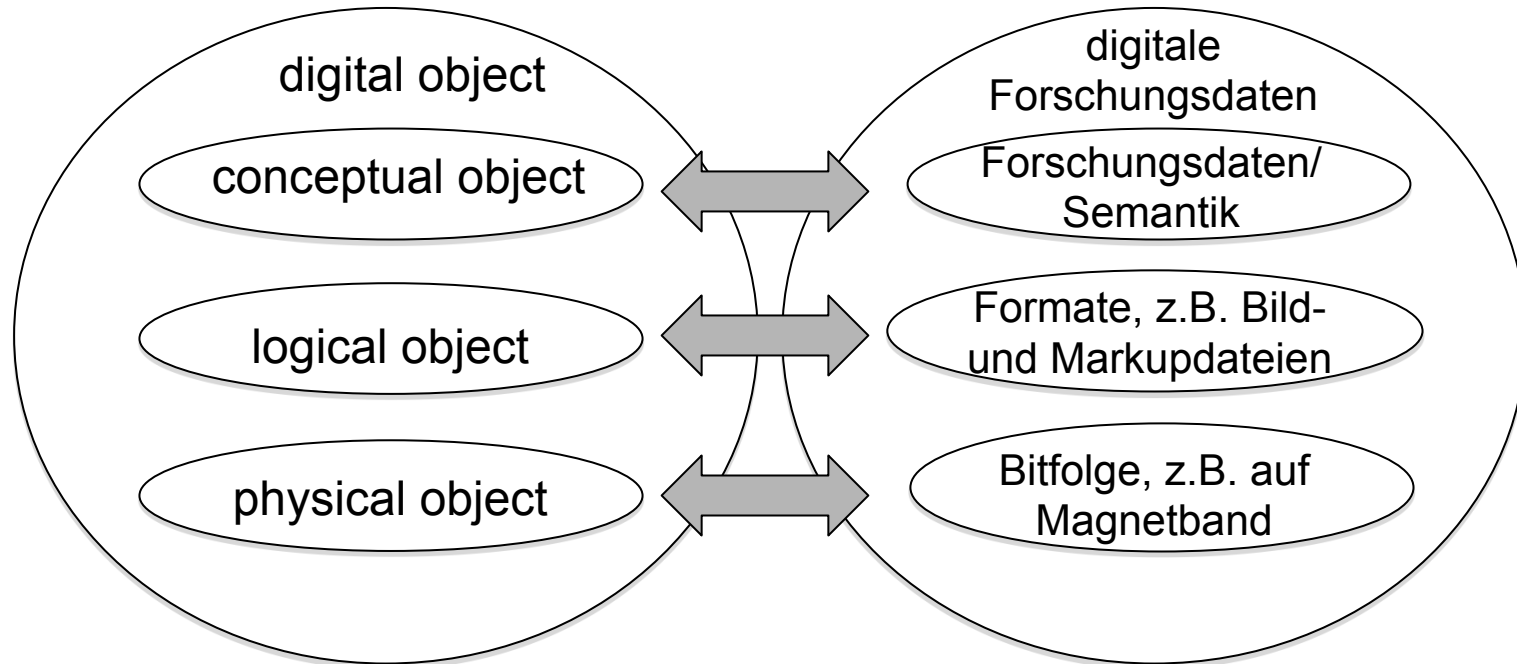


- Auftrag
- Vorgehen
- **LZA-Verständnis**
- Architektur
- Grid-Community-Unterstützung
- Weiterer Ablauf

- Bisher keine klare deutsche Terminologie
- „Langzeitarchivierung“ als Überbegriff für
 - das Sicherstellen der Nachnutzbarkeit
 - in einem anderen technischen, zeitlichen, fachlichen, organisatorischen oder sonstigen Kontext
- Dadurch große Schnittmenge mit aktuellen Begriffen wie Data Driven Science, Data Sharing, etc.
- „Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.“

Nature 461, 145 (10 September 2009), doi:
10.1038/461145a

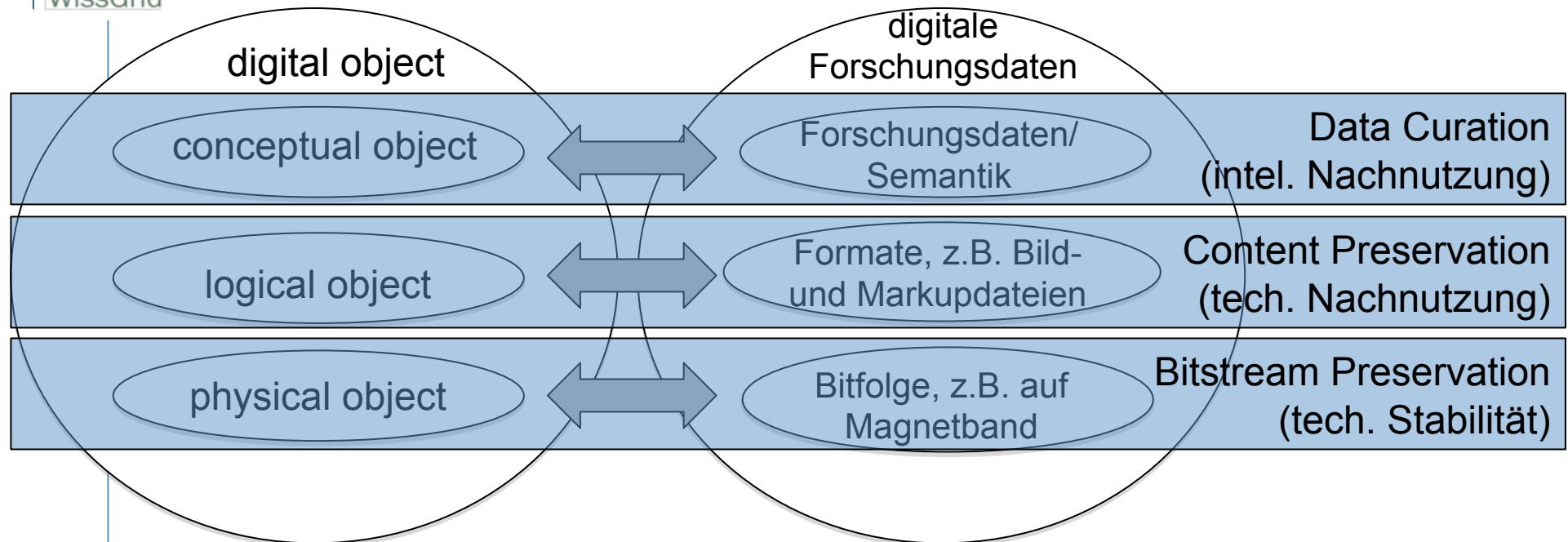




Angelehnt an Thibodeau: Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years, 2002. <http://www.clir.org/pubs/reports/pub107/thibodeau.html>



Drei Aspekte der Langzeitarchivierung

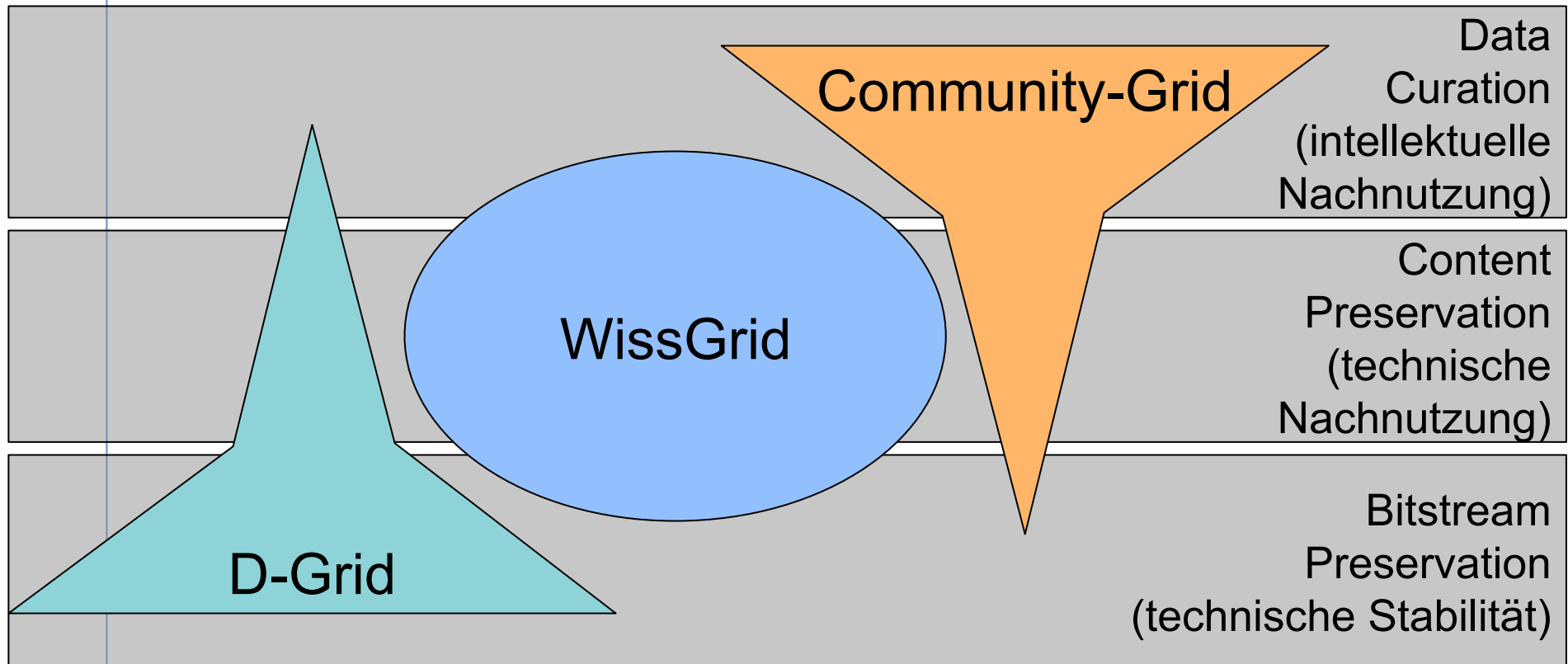


- **Data Curation: intellektuelle Nachnutzbarkeit**
 - Kontextinformationen, Objektmodelle, Versionierungen, ...
- **Content Preservation: technische Nachnutzbarkeit**
 - technische Qualitätskontrollen, Konvertierungen, ...
- **Bitstream Preservation: technische Stabilität**
 - genug unabhängige Kopien, Integritätsprüfung, ...



Kompetenzen für Langzeitarchivierung

- Für die Entwicklung von Infrastruktur und Werkzeugen zur LZA haben unterschiedliche Akteure die größte Kompetenz:





- Auftrag
- Vorgehen
- LZA-Verständnis
- **Architektur**
- Grid-Community-Unterstützung
- Weiterer Ablauf



- Rahmenbedingungen:
 - Kein einzelnes, technisches System
 - Keine/Kaum einheitliche, homogene Infrastruktur
 - technischer Wandel
 - große Community-Unterschiede und viele Spezifika
- Architektur ist
 - keine technische Detailarchitektur
 - sondern identifiziert wesentliche Funktionalitäten und Integrationsfaktoren
 - generisch, offen für neue Dienste, modular
 - wird ergänzt durch die Spezifikationen
- Organisatorische Faktoren werden in den LZA-Blaupausen behandelt werden



Analyse bisheriger LZA-Architekturen

- Analyse von internationalen Architekturen für LZA und Forschungsdaten
- Welche Funktionalität kann sinnvoll von wem entwickelt oder angeboten werden?
- Im Überblick...

Einzelne Dienste

Fachspezifische Dienste

Metadaten extraktion

Validierung, Qualitätskontrolle

Konvertierung

Archiv- und Speicherdienste

Preservation Planing

Integritäts-sicherung, Replikation

Repository: Ingest, Speichern, Access

Verzeichnisdienste

Rechte-management

Verzeichnisse für Schemata für Metadaten, Ontologien, ...

Suche

Repository: Metadaten-verwaltung

Infrastrukturdienste

Persistent Identifier Service

AAI, User Management

Monitoring, Logging

Workflow

Provenienz dienst

WissGrid

Community

D-Grid (DGI, Projekte, ...)

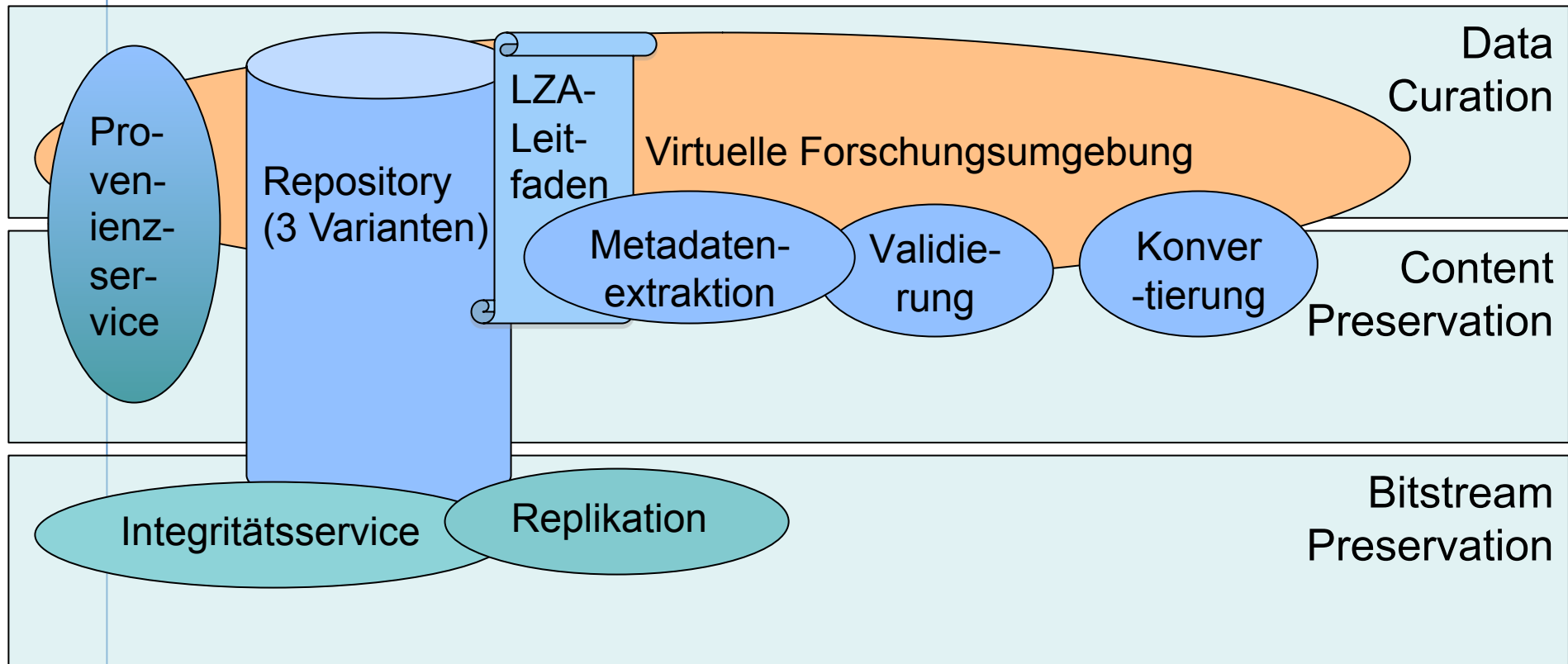


Übersicht der Entwicklungen

WissGrid =

Community =

D-Grid/Infrastrukturanbieter =





- Auftrag
- Vorgehen
- LZA-Verständnis
- Architektur
- **Grid-Community-Unterstützung**
- Weiterer Ablauf



Communities und Kooperationen

- Wie dargestellt eine Reihe von Aktivitäten um bedarfsgerecht zu entwickeln:
 - Begleitende AG mit Vertretern der Fachdisziplinen und SUN
 - Besuche bei Fachdisziplinen
 - Fallstudien im Architekturdokument, die fortgeschrieben werden
- Drei Community-Vertreter präsentieren später ihren Stand
- Die Rückmeldungen der Communities in einer Überblickstabelle:

	Bit Preservation (Verantwortung D-Grid)	Content Preservation						
		Metadaten-Extraktion	Validierung	Konvertierung	Provenance (Verantwortung D-Grid)	Repository		
						A	B	C
Astro	✗	○	○	○	○	✗	○	✗/○
C3	✗	○	✗ / ○	○	✗ / ○	✗	✗	✗
HEP	✗	—	—	—	—	✗	—	—
Medizin	✗	✗	○	○	○	○	—	—
Text	✗	✗ / ○	○	○	—	—	○	○
RadioAstro/ LOFAR	○	○	—	—	—	—	—	—
Klimafolgen	✗	○	○	○	○	✗	✗	✗
Photonen	○	○	○	○	○	○	○	○
Biostatistik	○	○	○	—	○	○	○	○
germ. Sprachwiss.	○	○/✗	○/✗	○/✗	○	—	○	—
Sowi (vorläufig)	○	○	○	○	○	—	○	○

Legende: ○ ... gewünscht

✗ ... vorhanden, Synergien möglich

Repository, Profil A: Grid-Workflowumgebung

Profil B: interaktive Umgebung

Profil C: Repositorien-Föderation



[Kooperation mit SUN]

- Sun, Forschung und Lehre
- Mitglied in der Community-AG
- Erste Gespräche haben gemeinsame Interessen gezeigt:
 - Bitstream Preservation: Anforderungen von WissGrid und Sun als Anbieter von Hardware und Management Software für heterogenen Storage
 - Data Curation: WissGrid liefert generische, zu integrierende Dienste und SUN ist Anbieter von Community-spezifischen Anpassungen
 - Technik: iRODS and Fedora spielen eine wichtige Rolle für WissGrid, Sun arbeitet an Forschungsumgebungen mit Fedora.



- Auftrag
- Vorgehen
- LZA-Verständnis
- Architektur
- Grid-Community-Unterstützung
- **Weiterer Ablauf**



10:30 – 11:15 **Spezifikation und Schnittstellen**

- **Forschungsdatenarchiv**
Andreas Aschenbrenner, SUB
- **LZA-Dienste**
Michael Lautenschlager, DKRZ

11:15 – 11:30 Pause

11:30 – 12:30 **Anwendungsfälle und Communities**

- **Photonenphysik**
Frank Schluenzen, DESY
- **germanistische Sprachwissenschaft**
Andreas Witt, IDS
- **Sozialwissenschaft**
Peter Bartelheimer, SOFI

12:30 – 13:00 **Diskussion**