



# Entwicklungen im Bereich Software-Architekturen und Software-Produkte langfristiger Informationserhalt und Grid Architekturen

Vascoda Informationsveranstaltung Fachliche Repositorien  
ZBW - Deutsche Zentralbibliothek für Weltwirtschaften - Kiel  
30.10.2007

Peter Rödиг

Peter.Roedig@unibw.de  
www.unibw.de/Peter.Roedig

**Fakultät für Informatik**  
**Institut für Softwaretechnologie**  
**Prof. Dr. Uwe M. Borghoff**  
**Tel.: 089/6004-2274 Fax: 089/6004-4609**  
**Uwe.Borghoff@unibw.de**  
**www.unibw.de/inf2/**

- Kurze Vorstellung
- Herausforderungen beim langfristigen Erhalt digitaler Information (LZA)
- Architekturen
  - Serviceorientierte Architekturen
  - Grid Architekturen
  - Anwendungsarchitekturen
- Gemeinsamkeiten zwischen LZA und Grid
- Anwendungsszenarien
- Entwicklungsstand

## Aktivitäten des Instituts für Softwaretechnologie mit Bezug LZA

### Grundlagen:

- Formalisierung der Migration (Erhalt relevanter Eigenschaften)
- Erweiterung / Verfeinerung OAIS-Referenzmodell

### Bücher:

- Langzeitarchivierung – Methoden zur Erhaltung digitaler Dokumente (dpunkt.verlag)
- Long-Term Preservation of Digital Documents. Principles and Practices (Springer)

### Mitarbeit in nestor-AGs:

- Vertrauenswürdige Archive – Zertifizierung
- Grid / eScience und Langzeitarchivierung
- LZA-Standards

### Projekte:

- Langzeitarchivierung digitaler Medien (Medienmigration) (DFG/BSB)
- Datenbankgestützte Langzeitarchivierung digitaler Objekte (DFG)
- Vergleich bestehender Archivierungssysteme (nestor)
- Mitentwicklung mediaTUM / integraTUM (Technische Universität München)
- Standards und Standardisierung im Kontext von Grid-Technologien und Langzeitarchivierung (nestor)

## Herausforderungen beim Langzeiterhalt digitaler Information

### 1. Inhärente Komplexität digitaler Objekte

- Logischer Abstand **physisches Medium** und **Information**  
„vom Datenträger über die Bitsequenz zur Information“
- Vielzahl möglicher interner (digitaler) Repräsentationsformen
  - Formatvielfalt
  - Virtualisierung
- Vielzahl möglicher externer Repräsentationsformen („Sichten“) durch (parametrisierbare) Operationen (Interaktion) auf den digitalen Objekten  
z.B.:
  - Navigieren in Textdokumenten
  - Datenbankabfragen
  - Durchschreiten virtueller Museen

## Herausforderungen beim Langzeiterhalt digitaler Information

### 2. Langfristigkeit

- Zeitlicher Abstand zwischen **Konsument von Information** und **Produzent von Information** (Verlust von Kontextwissen)  
z.B. Bedeutung von Symbolen zur Interaktion mit digitalen Objekten  
(Ablaufumgebung)
- Zeitlicher Abstand zwischen **Konsument von Information** und **Produzent / Nutzer von Werkzeugen** (Verlust von technischem Interpretationswissen + Umsetzung in Form von HW und SW)  
kurz: Technische Überalterung
- Integrität der physischen Medien

- Museumsansatz  
Aufbewahrung (Nachbau) und Betrieb der **Hardware** einschließlich Betriebssystem (= Computer-**Plattform**) plus Anwendungsprogramm (= Abspiel-/Ablauf-**Umgebung**)
- Emulation  
Nachbildung der **Hardware** oder der **Plattform** oder der **Umgebung** in Form von Software  
Digitale Objekte bleiben unverändert !  
*Portierung*: Migration der Software
- Migration  
Überbegriff für unterschiedliche „Transformationen“ digitaler Objekte  
Mehr oder weniger starker Eingriff in jedes Objekt !  
Schwer generalisierbarer Ansatz !

SOA:

Paradigma zur Organisation und Nutzung verteilter Fähigkeiten, die auch im Besitz unterschiedlicher Organisationsbereiche sein können

Kernkonzepte:

- **Sichtbarkeit:** gegenseitige Sichtbarkeit der Nachfrager und Anbieter von Fähigkeiten
- **Interaktion:** Aktivität zur Nutzung der Fähigkeiten
- **Real-World-Effekt:** Rückgabe von Information oder Zustandsänderungen an Entitäten, die an der Interaktion beteiligt sind
- **Service:** Mechanismus, um Nachfrage und Fähigkeit zusammenzubringen (umfasst Spezifikation der angebotenen Leistung und Angebot, die Leistung zu erbringen)

Höherwertigere Services durch Komposition von Services  
Inhaltliche Unabhängigkeit von konkreten Anwendungen

Bekannteste Ausprägung: Web Services (WSDL, SOAP, XML)

Grid:

Wandlung vom HPC zur Middleware und weiter zu einer Serviceorientierten Architektur [OGSA: Open Grid Services Architecture]

Kernkonzepte:

- Bündelung von Ressourcen (Pooling)
- Teilung von Ressourcen (Sharing)
- Virtualisierung von Ressourcen  
Abstraktion von konkreten (Implementierungs-)Eigenschaften
- Autonome (lokale) Verwaltung von Ressourcen

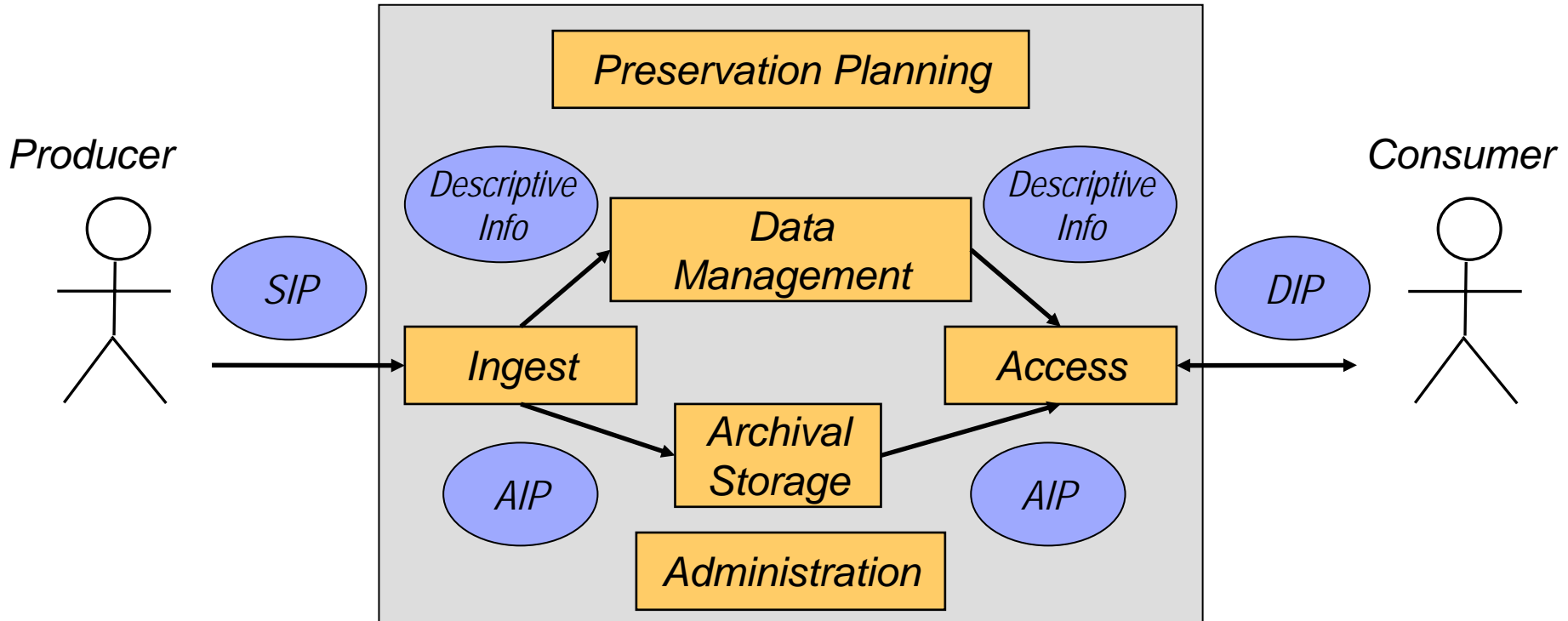
Dafür:

Definition einer Menge von Kernfähigkeiten (*Diensten*):

*Benennung von Objekten, Datendienste, Informationsdienste, Sicherheitsdienste, Job Management, Ressourcenmanagement, ...*

Ressourcen: CPU-Leistung, Speicher, Bandbreiten, aber auch Dienste

# Anwendungs-Architektur OAIS-Referenzmodell als Basis



IP (*Information Package*) = Daten + beschreibende Information für LZA

S: *Submission*

A: *Archival*

D: *Dissemination*

Hinweis: OAIS ist kein Design- oder Implementierungsmodell

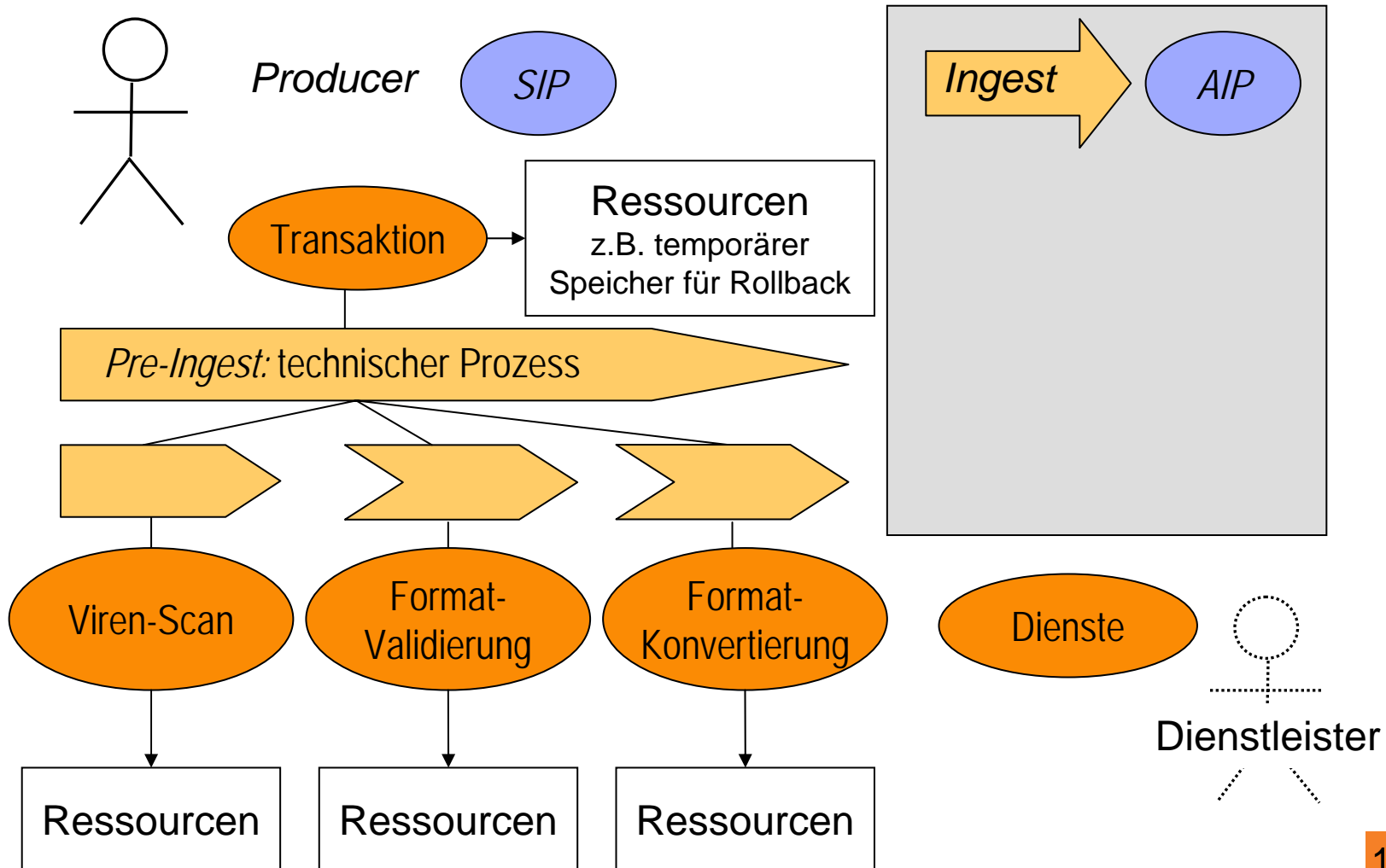
# Gemeinsamkeiten zwischen LZA und Grid

- **LZA:** als verteilte und kooperative Aufgabe
  - **Grids:** zielen ab auf die Integration, Virtualisierung und Verwaltung von Ressourcen und Diensten innerhalb verteilter, heterogener virtueller Organisationen [OGF]
- 

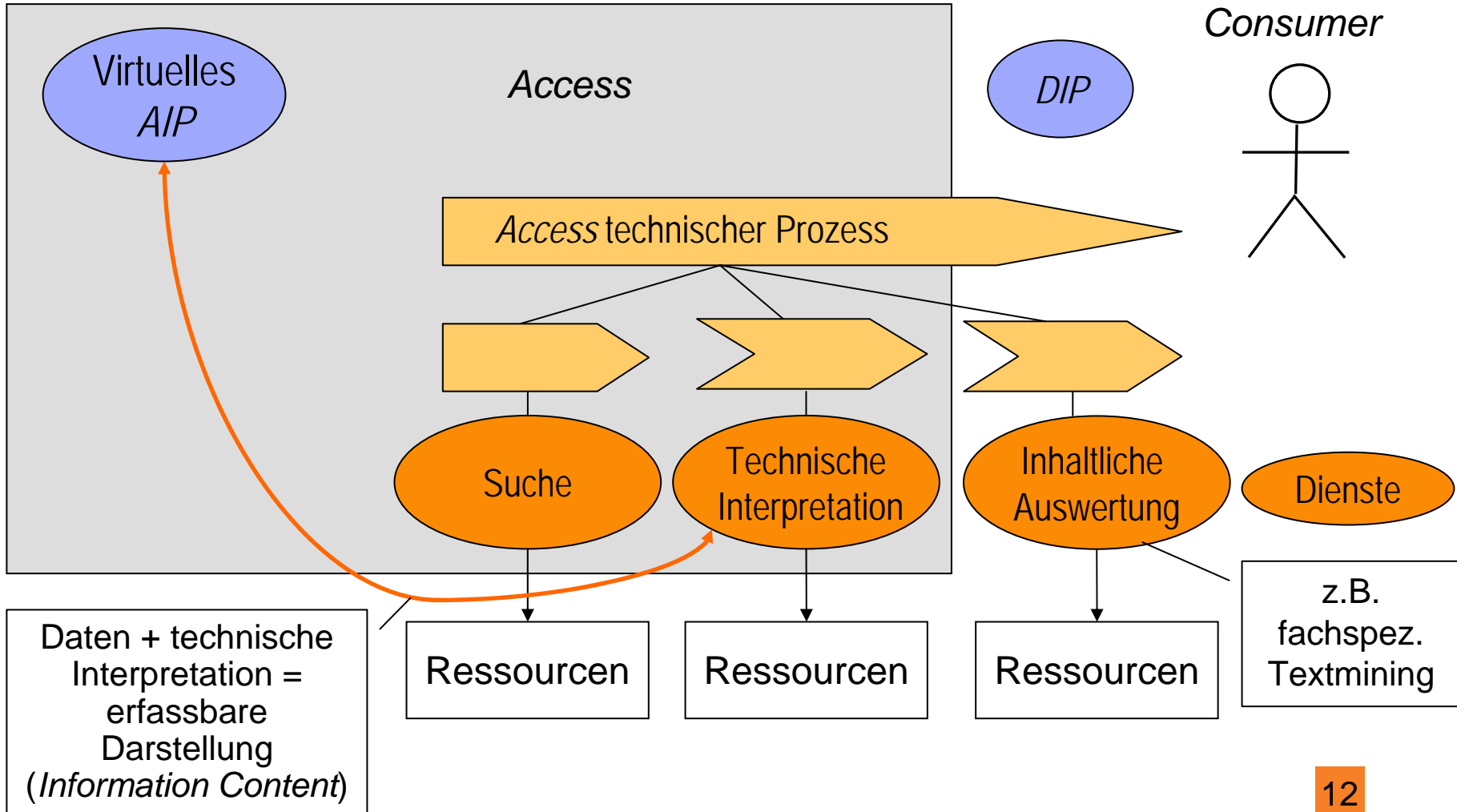
## Konzeptionelle und technische Anknüpfungspunkte zwischen LZA und Grids

- Finden, Identifizieren und Interpretieren von Daten
- Verwaltung und Publizieren von Diensten
- Behandlung heterogener Datenbestände („Messy Data“)
- Protokollierung von Bearbeitungsschritten („Provenancing“)
- Identitäts- und Rechtemanagement
- Abstraktion von technischen Eigenschaften  
(Migrationsfähigkeit der technischen Infrastruktur)
- Sichere und konsistente Speicherung von Daten  
Datenreplikation und Synchronisation
- Qualitätssicherung beim *Ingest*: Formatvalidierung, Formatkonvertierung  
(Inanspruchnahme Rechenleistung)
- Unterstützung beim *Access*: Interpretation, Mining  
(Inanspruchnahme Rechenleistung)

# Grid-Dienste im LZA-Kontext: Szenario 1



# Grid-Dienste im LZA-Kontext: Szenario 2



### Langfristige Vertrauenswürdigkeit ???

Zuverlässigkeit und Verfügbarkeit von Diensten

- Authentifizierung, Autorisierung
- Auditing (Provenance)
- Auffindbarkeit und Identifizierung von Objekten
- Materialisierbarkeit virtueller Datenobjekte
- Interpretierbarkeit von Bit-Sequenzen  
zur Erstellung von Informationsobjekten

Voraussetzungen für Vertrauenswürdigkeit

- Standardisierung
- Organisatorische Nachhaltigkeit (Grenzen des Selbstmanagements)

# Stand der Entwicklung grober Überblick

## SOA:

- Anwendungen: in (LZA-)Standards und (LZA-)Produkten tw. zu finden
  - METS: Verknüpfung von digitalen Objekten mit Diensten
  - Produkte: FEDORA, DSpace, ... , sowie Basisprodukte wie Datenbanksysteme
- Standardisierung:  
immer noch in einer lebhaften Phase

Keine Definition von Domänen-spezifischen Lösungen (auch nicht LZA) / Basis-Technologie

## Grid:

- Anwendungen:  
Produkte verfügbar, Handhabung und Integration entwicklungsbedürftig, weitere Grundlagenarbeit und Testbeds erforderlich, LZA-Aspekte bisher untergeordnet
- Standardisierung:  
eher am Anfang, Notwendigkeit erkannt, Abstützung auf vorhandene Standards (W3C, IETF, OASIS, ...)

Zur Vertiefung s. nestor-AG „Grid / eScience und LZA“, insbesondere Expertisen (Anfang 2008):

- Anforderungen von eScience und Grid-Technologie an die Archivierung wissenschaftlicher Daten  
- Geoforschungszentrum Potsdam
- Synergiepotentiale zwischen GRID- und eScience-Technologien für die Langzeitarchivierung  
- FernUniversität Hagen

Vielen Dank für Ihre Aufmerksamkeit!



### Nachricht an die Buchbinder.

Weiln auf diesen Tabellen weder Pagina, Signatur noch Custos vorhanden, die Umstände auch solche anzubringen nicht erlaubt; Als hat man hiermit erinnern wollen, daß allezeit die gerade Tabelle, als: II. IV. VI. VIII. X. u. s. f. in die Mitten hinein gepfalzet werden muß, alsdenn das ganze Werk in gehörige Ordnung kommen wird.

HPC	High Performance Computing
OAIS	Open Archival Information System -- Reference Model, ISO 14721:2003
OASIS	Organization for the Advancement of Structured Information Standards
OGF	Open Grid Forum
IETF	Internet Engineering Task Force
METS	Metadata Encoding and Transmission Standard
SOAP	Protokoll zum XML-basierten Austausch von Nachrichten, häufig auf Basis von HTTP/HTTPS (ursprünglich für Simple Object Access Protocol)
WSDL	Web Services Description Language
Transaktion	Folge von Operationen, die entweder komplett oder gar nicht durchgeführt wird (Rollback im Fehlerfall)
Virtuelles AIP	In Szenario 2 wird aus den Daten mit Hilfe technischer Interpretation zum Abfragezeitpunkt ein Objekt zusammengebaut, das für den Endnutzer (Consumer) Information darstellt. Dieser Vorgang kann z.B. die Bereitstellung eines geeigneten Emulators umfassen, der ein obsoletes Format in der jeweils aktuellen Umgebung anzeigen kann. Bleibt der Vorgang für den Endnutzer verborgen, erscheint es, als sei tatsächlich ein fertiges Informationsobjekt vorhanden.
Zu nestor	<a href="http://www.langzeitarchivierung.de">www.langzeitarchivierung.de</a>