eSciDoc – Stand und Ausblick

Leni Helmes
Matthias Razum

Fachinformationszentrum Karlsruhe

Oktober 2006

FIZ Karlsruhe

Das eSciDoc-Projekt

- eSciDoc – ein vom BMBF gefördertes gemeinsames Projekt der Max-Planck-Gesellschaft und FIZ Karlsruhe zum Aufbau einer
  - integrierten Informations-, Kommunikations- und Publikationsinfrastruktur
  - für netzbasiertes wissenschaftliches Arbeiten
  - am Beispiel von multidisziplinären Anwendungen in der Max-Planck-Gesellschaft (MPG)
- MPG verantwortlich für Anforderungen und Applikationen
- FIZ Karlsruhe verantwortlich für Framework, Nachhaltigkeit und Nachnutzung durch andere Forschungseinrichtungen
- Mitgestaltung von e-Science

e-Science erfordert Infrastruktur

- *“e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it” .* John Taylor (Director General of Research Council, UK)

- Die Abkürzung e-Science steht für "enhanced science" – also für eine in ihrer Arbeitseffizienz durch neue Technologien gesteigerte Wissenschaft.
  - Die Verbesserungen eröffnen sich aus den technischen Möglichkeiten der nächsten Generation der Informationstechnologie und des Internets (Web 2.0).

Neue Formen wissenschaftlichen Arbeitens

- Nutzer = Informations**konsument** + Informations**produzent**
- Information = Netzwerk aus Digitalen Objekten
  - die auf Anfrage dynamisch zusammengestellt und weiter verarbeitet werden können
  - und somit nicht immer ein ‚wirkliches' Analogon haben (wie z.B. ein Zeitschriftenartikel)

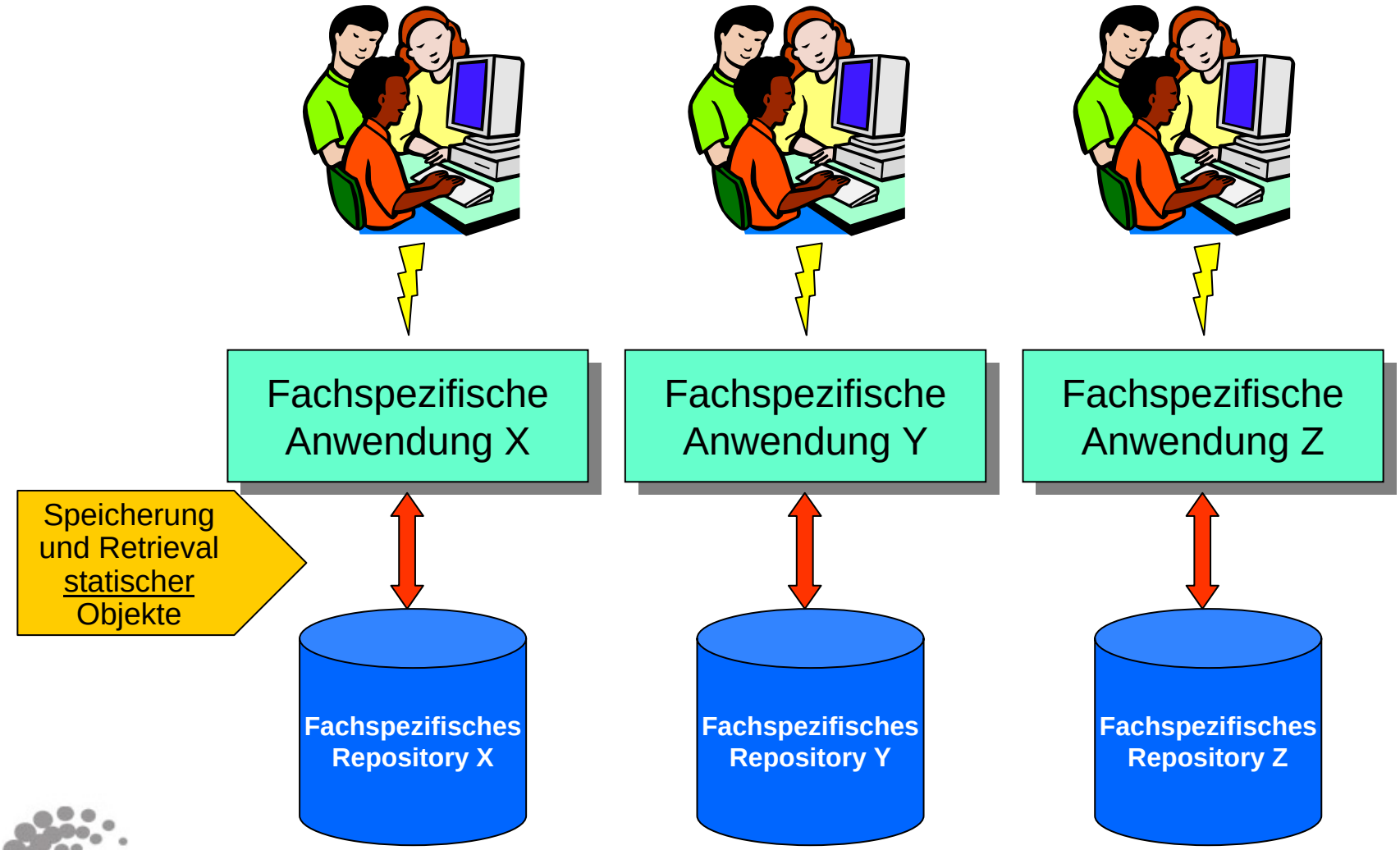
- Daraus resultieren neue Formen des wissenschaftlichen Arbeitens
- Diese erfordern innovative Informationsinfrastrukturen und Dienstleistungen

Dynamische Digitale Objekte

- Diese Digitalen Objekte (DO) sind nicht mehr nur statisch
- Sie "durchleben" einen Workflow, in dessen Verlauf sie von vielen Akteuren (Personen, Software, Instrumente) angefasst, verändert und mit anderen zu neuen DOs kombiniert werden
- Dies wirft vollkommen neuartige Fragestellungen im Sinne ***Guter Wissenschaftlicher Praxis*** auf, z.B.:
  - Wem darf das Objekt zu welchem Zweck "ausgeliefert" werden? (Rollen und Zugriffsrechte)
  - Wie viel vom Objekt-Lebenslauf muss eindeutig nachvollziehbar sein und langfristig erhalten werden (Zitierfähigkeit, stabile Links)

➢ Innovative und nachhaltige Konzepte – nicht nur technische - sind gefragt!

e-Science erfordert Infrastruktur – keine geschlossenen und statischen Systeme …



e-Science erfordert Infrastruktur – disziplinübergreifend, standardisiert und offen

Scholarly Workbench

Publication Management

Framework

Prozesssteuerung

Authentifizierung & Rechteverwaltung

Suche

Interdisziplinäres Repository "Gedächtnis der MPG"
(Primärdaten, Metadaten, Texte, Bilder, Videos, Software, Simulationen, Audiodateien ...)

Verwaltung von Organisationen und Personen

Metadaten-verwaltung

Format-verwaltung

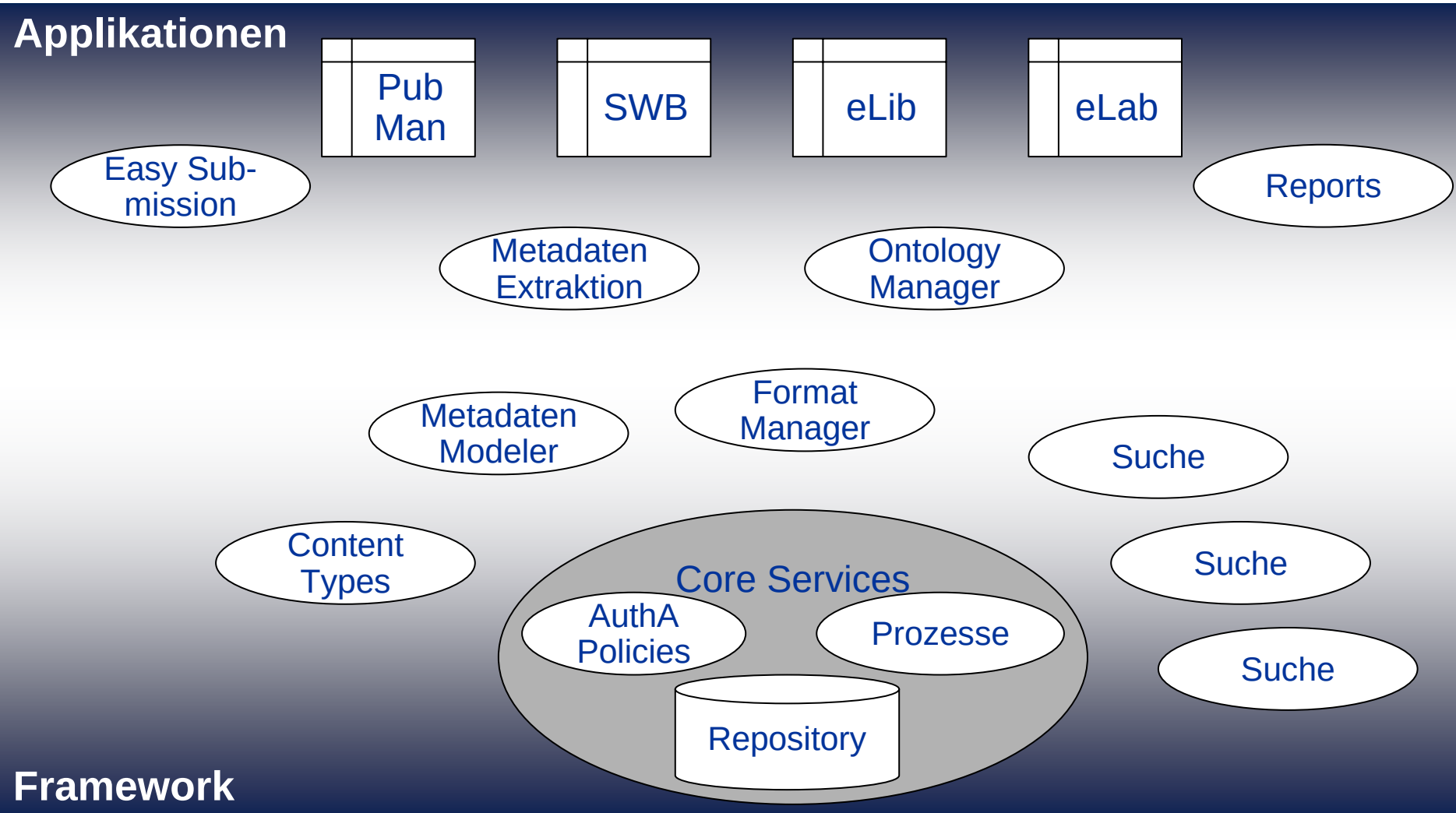
Import, Export

Interaktives Arbeiten mit den ‚lebenden' Daten im Repository

eLib

eLab Journal

eSciDoc: Framework plus Applikationen



eSciDoc-Framework

- Das Framework stellt alle notwendigen Funktionen zur Verfügung, um den gesamten Lebenszyklus eines wissenschaftlichen Dokuments sowie den zu seinem Verständnis notwendigen Kontext abzubilden.
- So kann bereits der erste Entwurf einer wissenschaftlichen Arbeit im Framework abgelegt und mittels entsprechender Berechtigungsmechanismen nur einem begrenzten Auditorium zugänglich gemacht werden.
- Dabei gibt es nahezu keinen Beschränkungen für den Aufbau und den digitalen Bestandteilen eines wissenschaftlichen Dokuments.

Komponenten des eSciDoc-Frameworks

Zentrale Komponenten

- Repository
- Prozesssteuerung
- Authentifizierung und Rechteverwaltung

Mehrwertdienste

- Verwaltung von Organisationen und Personen
- Metadatenverwaltung
- Formatverwaltung
- Import und Export
- Suche

eSciDoc Repository /1

Das Kernstück einer jeden e-Science-Anwendung ist das Repository.

Dieses

- speichert die digitalen Objekte
- legt bei Änderungen Versionen an und
- verknüpft die Objekte mit Verwaltungsdaten (z.B. beschreibende, technische und administrative Metadaten, Konfigurationen, Änderungshistorie).

eSciDoc Repository /2

Anwendungsneutrale Speicherung

- ermöglicht die Wiederverwendung von Objekten in neuen Kontexten und steigert damit deren “Wert”
- erleichtert die Archivierung

- Das bewährte Open Source-Projekt Fedora (Flexible Extensible Digital Object Repository Architecture) bildet die Grundlage des Repositories.
- Durch eine von den momentanen Nutzungsszenarien unabhängige Modellierung der Objekte in Fedora können diese später in neuen, bisher noch nicht angedachten Anwendungskontexten wieder verwendet werden.
- So soll das Ziel erreicht werden, mit einem Repository eine Vielzahl von Anwendungen dauerhaft 'speisen' zu können.

eSciDoc Repository /3

Versionierung

- Verwaltung des gesamten Object Lifecycles
- sichert korrekte Zitierung auch bei sich dynamisch entwickelnden Objekten / *Good Scientific Practice*
- Berücksichtigt auch Aggregationen (z.B. Collections, Bundles) und ermöglicht so digitale Editionen

eSciDoc Repository /4

Beispiel für Versionierung:

- Ein Autor lädt einen Artikel als Preprint ins Repository, nimmt ihn bereits zu diesem Zeitpunkt in seine persönliche Publikationsliste (die er über das Publication Management automatisch erzeugen lassen kann) auf – somit veröffentlich er den Artikel. Damit erhält der Artikel einen Persistent Identifier. Leser können darüber den Preprint zitieren und referenzieren.
- Später überarbeitet der Autor seinen Artikel nochmals für die Publikation in einer Zeitschrift. Auch diese Version liegt im Repository. Damit Leser stets wissen, auf welche Version sich ein Zitat oder eine Referenz bezieht, müssen sich beide Versionen mit unterschiedlichen Persistent Identifiern im System finden lassen. Weiterhin sollte es möglich sein, zu späteren und früheren Versionen zu springen, also eine Versionshistorie anzeigen zu können.

eSciDoc Repository /5

Unterstützung von Objektrelationen

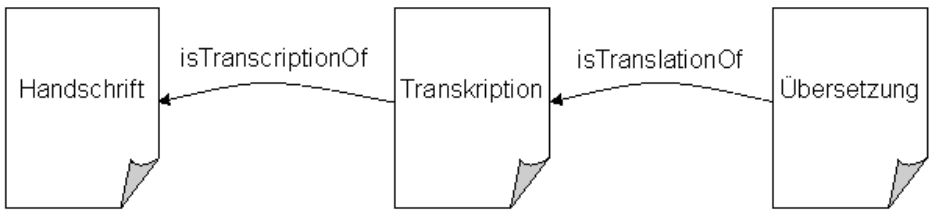
Diese

- werden durch RDF ausgedrückt
- sind über einen Triple Store (Kowari) indexiert und suchbar
- ermöglichen die Einführung z.B. eines Ontologie-Managers
- auch Verweise auf externe Ressourcen (z.B. Primärdaten) sind möglich

eSciDoc Repository /6

Beispiel für Objektrelationen:

- Eine Autorin reichert die digitalisierte Form einer alten lateinischen Handschrift an. Dazu verfasst sie eine Transkription der Handschrift. Somit liegt der Text maschinenlesbar vor und kann beispielsweise im Volltext durchsucht werden.
- Später fügt ein Kollege eine Übersetzung in Deutsch hinzu. Die digitalisierte Handschrift, ihre Transkription und Übersetzen stehen untereinander in Beziehung. Genau diese Beziehungen lassen sich über RDF ausdrücken:



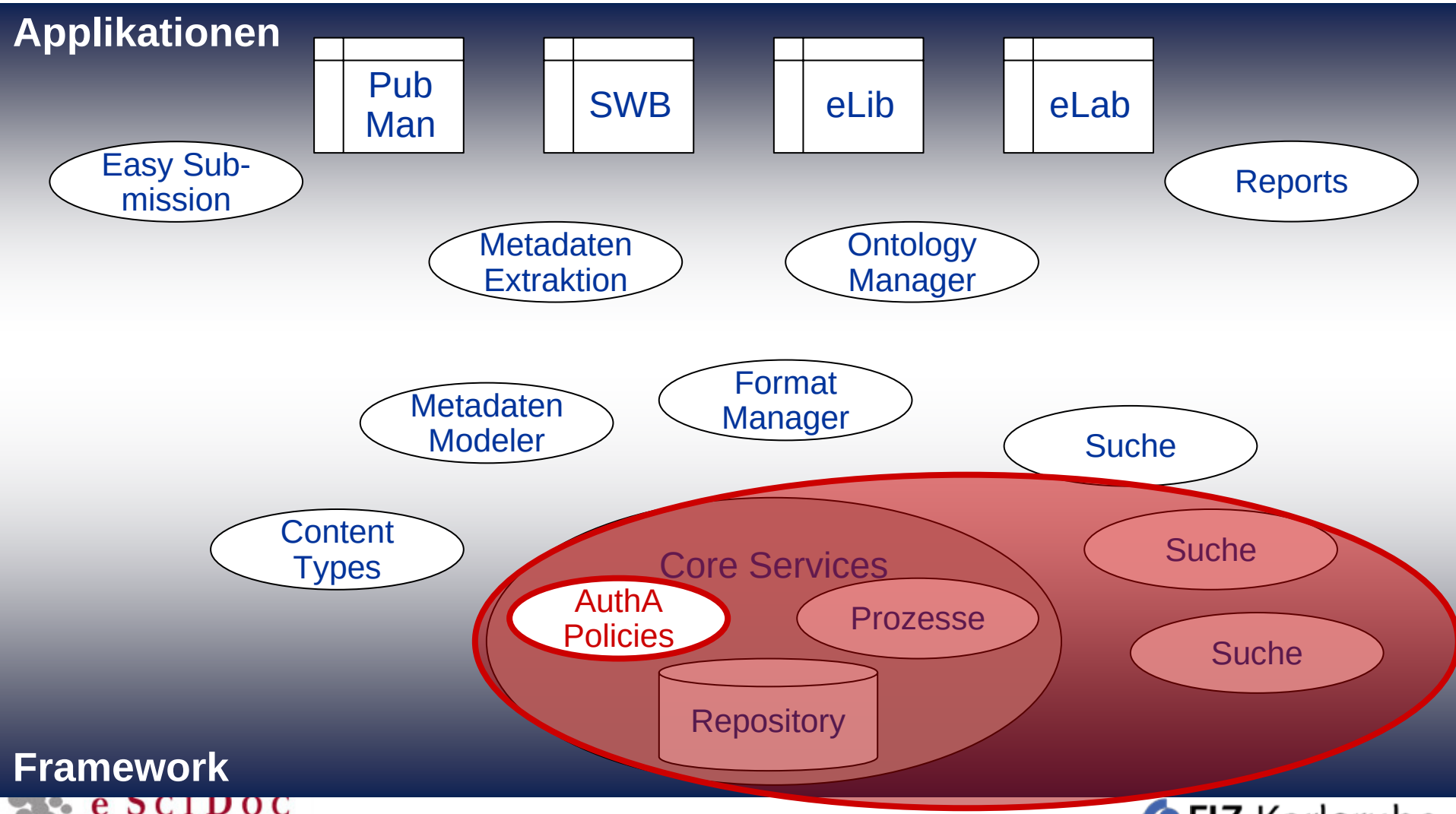
eSciDoc Prozesssteuerung

- Flexible Workflows, z.B. für Submission, Ingestion und Withdrawal
- Einzelne Prozessschritte werden als Webservices implementiert
  - ermöglicht Wiederverwendbarkeit
  - können auch außerhalb von eSciDoc realisiert sein
- Kombination von manuell durchzuführenden und automatisch ablaufenden Schritten möglich

eSciDoc Authentifizierung und Rechteverwaltung /1

- Komplexe Zugriffsrechteverwaltung auf Basis von
  - Benutzern
  - Rollen
  - Aktionen
  - Objekten
- Kann für Services durch Konfiguration (AOP) einfach ein- und ausgeschaltet werden
- Trennung von User Management und Rechteverwaltung, um später verteilte Authentifizierung (Shibboleth, Liberty) zu ermöglichen

eSciDoc Authentifizierung und Rechteverwaltung /2



Verwaltung von Organisationen und Personen

- Digitale Objekte haben Beziehungen zu Personen, z.B. als Autoren oder Herausgeber.
- Diese wiederum haben eine Zugehörigkeit mit einer Institution und ihren Organisationseinheiten wie Abteilungen oder Arbeitsgruppen.
- Das eSciDoc-Framework stellt Dienste zur Darstellung dieser Zusammenhänge bereit, um darüber z. B. Eigentums- und Urheberverhältnisse darzustellen und indirekt auch die Zugriffsrechte der Objekte zu steuern.
- Weiterhin sind über diese Dienste eine Zuordnung von Personen und Organisationseinheiten auf die in der Benutzerverwaltung bereitgestellten Daten möglich.

eSciDoc Metadatenverwaltung

- Unterstützt das Mapping von Metadatenschemata aufeinander
  - Abbildung externer MD-Schemata auf das eSciDoc-Schema (Ingestion)
  - Transformation des internen Schemas auf externe (Export, Harvesting)
- Bietet Werkzeuge zum Erstellen neuer Abbildungsregeln
- Validiert Metadaten anhand hinterlegter XML-Schemata

eSciDoc Formatverwaltung

- eSciDoc unterscheidet drei Klassen von Dateiformaten:
  - intern – bekanntes Format, das vom System interpretiert werden kann und sich auch für Archivierung eignet
  - bekannt – bekanntes Format, für das eine Transformation in ein internes Format möglich ist. Es werden sowohl das bekannte wie das interne Format im System gespeichert
  - unbekannt – kann vom System nur als Binary Stream gespeichert und nicht umgewandelt oder interpretiert werden.
- Der Format Manager bestimmt die Formate, extrahiert technische Metadaten und führt gegebenenfalls die Konvertierung durch

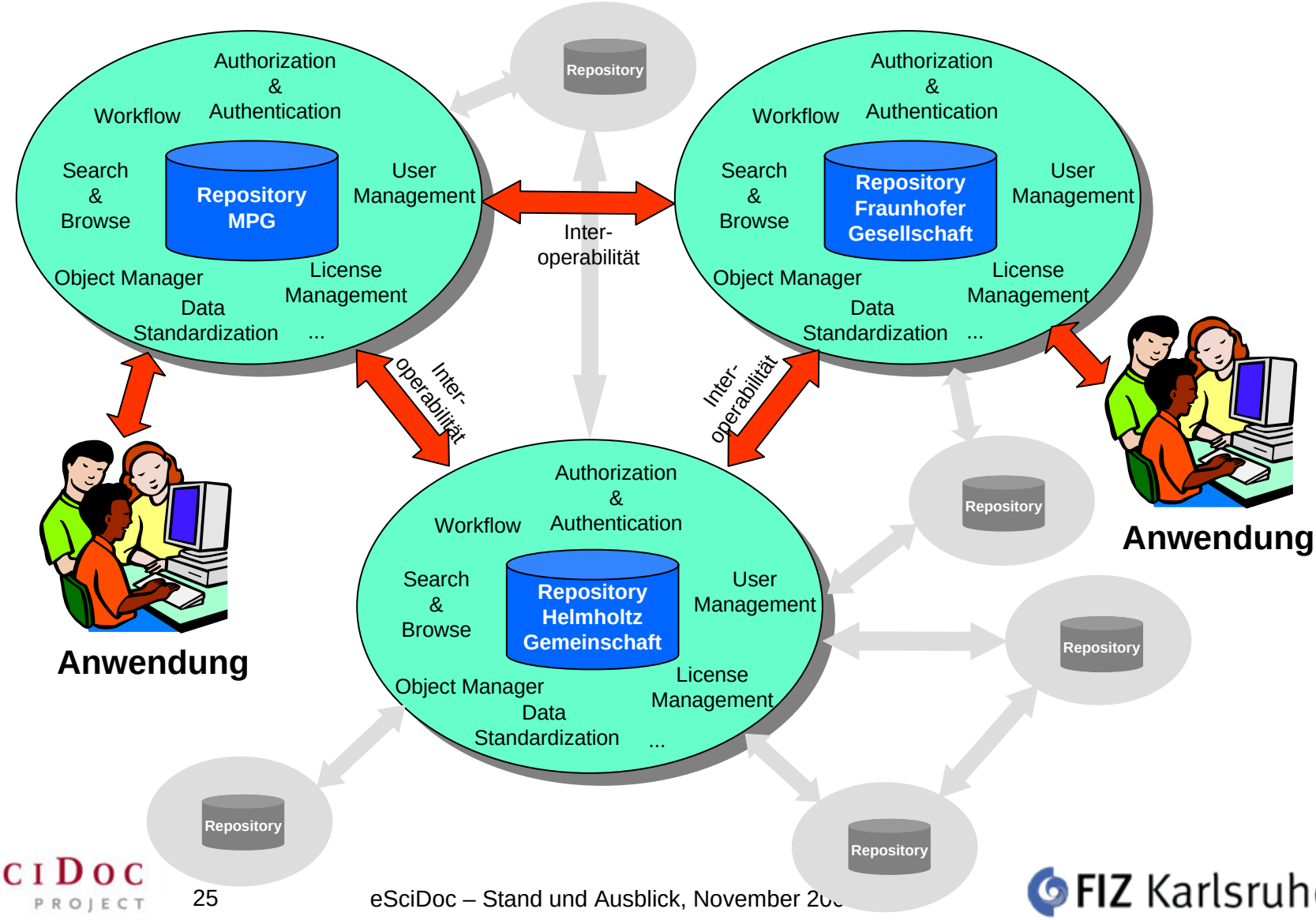
Import und Export

- Flexible Möglichkeiten zum Datenaustausch mit anderen Systemen.
- Flexibilität nicht nur hinsichtlich der Formate, sondern auch der Automatisierung solcher Prozesse.
- Framework-Dienste wie Prozesssteuerung, Format- und Metadatenverwaltung spielen hier eine wichtige Rolle
- Unterstützung gängiger Formate und Standards (z.B. OAI-PMH, Endnote, BibTex, …)

eSciDoc Suche

- Neben einer allgemeinen Suche auf Basis der Volltexte sowie des eSciDoc-internen Metadatensatzes können weitere, spezialisierte Suchen aufgesetzt werden
- Suchen werden als Service außerhalb des eSciDoc-Kerns aufgesetzt und können auch von Instituten oder Wissenschaftlern direkt implementiert werden
- Über eine flexible Architektur für die Indexierung können auch komplexe Anforderungen im Bereich morpholgischer Analysen (Normalisierung, Lemmatisierung, Normdateien) umgesetzt werden

Vision: eSciDoc - Nukleus der e-Infrastruktur für die Wissenschaft



Aktueller Stand

- Die Konzeptphase ist weitgehend abgeschlossen
- Die Implementierung des Frameworks und der Applikationen Scholarly Workbench und Publication Management hat begonnen
- Enge Einbindung der Anwender innerhalb der MPG (Piloten)

Ausblick

- Erste Ergebnisse Anfang 2007
- Demonstrationen der Applikationen PubMan und Scholarly Workbench zur German e-Science Konferenz im Mai 2007
- Erste öffentliche (innerhalb der MPG) Releases Mitte 2007

- Weitere Services, aufbauend auf dem Framework, im Verlauf des Jahres 2007 - aus dem Kreis der Piloten und auch von Dritten

Nachnutzung

- Sowohl Applikationen wie auch Framework stehen zur Nachnutzung zur Verfügung
- Gespräche mit Interessenten aus anderen Wissenschaftsorganisationen finden bereits statt
- Wichtig: Abstimmung mit anderen Projekten, z.B:
  - DRIVER
  - links4science
  - Fresco
  - TextGRID
  - DGrid
  - kopal

Danke!

FIZ Karlsruhe