eSciDoc – Das Repository-Konzept der Max Planck Digital Library

Malte Dreyer | malte.dreyer@mpdl.de
Ulla Tschida | tschida@mpg.de

Basierend auf der eSciDoc Infrastruktur entwickelt die MPDL ein Set an Services und Anwendungen, die als Repository für unterschiedliche Forschungsdaten, u.a. Publikationsdaten, in der MPG eingesetzt werden. Der Beitrag skizziert ein möglichst umfassendes Bild der Ausgangslage, der fachlichen und technischen Anforderungen sowie des gewählten Entwicklungsansatzes. Zusätzlich wurde ein Schwerpunkt auf die organisatorischen und sozialen Aspekte eines Infrastruktur-Projektes für eScience in einer großen Wissenschaftsorganisation gelegt.

Einleitung

eSciDoc ist ein vom BMBF gefördertes gemeinsames Projekt der Max-Planck-Gesellschaft (MPG)[1] und dem Fachinformationszentrum FIZ Karlsruhe[2]. Innerhalb der MPG ist das Projekt an der zum 1. Januar 2007 gegründeten Max Planck Digital Library (MPDL)[3] angesiedelt. Das Ziel des Projekts ist die Entwicklung einer disziplinübergreifenden virtuellen Forschungsumgebung im Rahmen der eScience-Initiative des Bundes. Die Infrastruktur und darauf aufbauende Anwendungen stehen unter *www.escidoc.org* im Rahmen einer Open-Source-Lizenz zur Verfügung. Innerhalb der MPG werden die entwickelten Anwendungen zurzeit für den Produktiveinsatz vorbereitet, außerhalb der MPG evaluieren international über 20 Einrichtungen einen möglichen Einsatz oder entwickeln bereits erste eigene Anwendungen und Services. Die weitere Entwicklung und Pflege von Infrastruktur und Anwendungen wird auch über das Ende der Förderung zum 31. Juli 2009 hinaus durch eigene Kapazitäten von MPG und FIZ Karlsruhe sichergestellt.

Kernanwendungen, die auf der Infrastruktur aufsetzen, sind zurzeit ein System zum Publikationsdatenmanagement „PubMan“, ein System zur kollaborativen Arbeit mit Bild-Daten „FACES“ sowie ein System zur Verwaltung und Anreicherung von digitalisierten Text-Ressourcen „ViRR“.

1 Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V., http://www.mpg.de
2 Fachinformationszentrum FIZ Karlsruhe, http://www.fiz-karlsruhe.de
3 Max Planck Digital Library, http://mpdl.mpg.de

Ausgangslage – eScience-Herausforderungen

Heterogenität der Ressourcen

Durch die zunehmende Datenzentriertheit der Forschungsaktivitäten steigt der Bedarf an technischen Infrastrukturen, die unterschiedliche digitale Ressourcen und Artefakte der wissenschaftlichen Arbeit transparent bearbeitbar und nachhaltig verfügbar halten. Für diesen Zweck werden zurzeit meist einzelne dedizierte Anwendungen mit speziell entwickelter Datenhaltung entwickelt. Zunehmend virtuelle, globale und interdisziplinäre Forschungsansätze erfordern jedoch die Identifikation und Zugänglichkeit von Ressourcen unterschiedlicher Herkunft sowie Möglichkeiten zur Anreicherung, zur Verbindung und Integration von Ressourcen und die Anwendung von disziplinspezifischen Werkzeugen, um die Ressourcen in neuen Kontexten zu nutzen.

Technologiegestützte Kommunikation

Kommunikation von Ergebnissen ist eine grundlegende Bedingung für wissenschaftlichen Fortschritt. Wurden bislang die Ergebnisse selektiv und aggregiert in Form einer wissenschaftlichen Publikation veröffentlicht, ermöglichen zunehmend webbasierte Forschungsaktivitäten eine transparente und durchgehende Dokumentation der einzelnen Arbeitsschritte, z. B. auf welcher Grundlage gearbeitet wurde (i. e. Qualität der verfügbaren digitalen Ressourcen bzw. Entitäten), wie die Ausgangslage bearbeitet oder modifiziert wurde (i. e. Qualität der Vorgehensweisen

und Methoden) und welche (Zwischen-) Ergebnisse für die weitere Forschung verwandt wurden[4]. Die bewusste Zugänglichmachung der wissenschaftlichen Artefakte und ihrer jeweiligen Kontexte ermöglicht somit neuartige und vor allem unerwartete Ansetzungspunkte für neue Forschungsfragen. Ressourcen sind somit ebenso einem stetigen Änderungsprozess unterworfen wie auch die entsprechenden begleitenden Informationen zur Herkunft, zur inhaltlichen Beschreibung oder zu den Nutzungskontexten. eScience-Anwendungen unterstützen diese kommunikativen Prozesse und müssen deshalb als Arbeitsumgebungen wissenschaftszentriert konzipiert und gestaltet werden. Der Begriff „Usability" erhält somit eine umfassende Bedeutung als wissenschaftsnahe Gestaltung von Arbeitsabläufen zur Suche und Identifikation relevanter Wissensbestände im Netz, zur transparenten Beschreibung von angewandten Ressourcen, Methoden und Ergebnissen sowie zum Einsatz kooperativer Werkzeuge zum Austausch von Artefakte und über Artefakte.

Semantic Web

Begleitend zu diesen Entwicklungen stehen die informationswissenschaftlichen Konzepte des „Semantic Web", um Ressourcen für diese Vorhaben günstig aufzubereiten und angereichert bereitzustellen. „The coolest thing to do with your data will be thought of by someone else"[5] umschreibt das Potential sowie die Herausforderung an maschinenlesbare Semantik von Daten, ihren Kontexten und Inhalten. Technologien wie Ontologien, Collection-, Service- oder Metadaten-Registraturen können hier die Identifikation und Nachnutzbarkeit von relevanten Entitäten wie Daten, Dokumenten, Personen, Organisationen oder Konzepten erhöhen.

4 Nicht nur innerhalb der wissenschaftlichen Community, sondern verstärkt im öffentlichen und politischen Kontext, sind die Forderungen nach Kommunikation von wissenschaftlichen Ergebnissen spürbar: Förderorganisationen, Industrie, Nicht-Regierungsorganisationen, Politik, die Zivilgesellschaft und andere soziale Gruppen erwarten eine transparente Kommunikation von wissenschaftlichen Tätigkeit und Ergebnissen.

5 Leitspruch der JISC Common Repository Interfaces Group CRIG http://www.ukoln.ac.uk/repositories/digirep/index/CRIG

Technische Anforderungen

Um diese nur grob skizzierte Ausgangslage zu adressieren, wurden in der Entwicklung der eSciDoc-Infrastruktur verschiedene technische Anforderungen a priori definiert, die im Laufe der Entwicklung einzelner Anwendungen konsolidiert werden.

Entitäten, Versionierung

Die Verwaltung von digitalen Ressourcen beinhaltet nicht nur die reine Ablage von Bits und Bytes, sondern auch die entsprechenden Zusatzinformationen bzw. Metadaten in ihrem Bezug zu den Ressourcen, die zu einem beliebigen Zeitpunkt nachvollziehbar abgebildet werden müssen. Diese komplexeren Anforderungen werden durch entsprechende Versionierungssysteme, wie z. B. „Subversion", nicht direkt unterstützt und erfordern eine breitere Sicht auf mögliche Informationseinheiten, ihre jeweiligen Komponenten und Metadaten.

Zugriffsrechte

Darüber hinaus sind im Kontext von Forschungsaktivitäten Ressourcen teilweise frei zugänglich, teilweise mit sehr eng gefassten Zugriffsrechten verbunden, wobei sich die Zugriffsrechte auch während eines Lebenszyklus verändern können. Um vielfältige Szenarien möglichst feingranular, auch in dezentralen Authentifizierungsmodellen, abbilden zu können, müssen entsprechende Technologien unterstützt werden.

Interoperabilität

Zum Austausch von Daten mit anderen Systemen, zur Anbindung bestehender Anwendungen oder disziplinspezifischer Werkzeuge ist ein breites Spektrum an disziplin- oder technologiespezifischen Schnittstellen sowie die Verwendung von standardisierten Protokollen und offenen Formaten erforderlich. Ebenso sind vielfältige Sichten und Präsentationsformen von Ressourcen und deren Zusammenstellungen zu unterstützen und in der Infrastruktur formal als Content Models zu beschreiben.

Content Models

Die Quantität und Heterogenität wissenschaftlich relevanter Ressourcen erfordert die Möglichkeit, auch grundsätzlich unbekannte Inhalte und entsprechende Metadaten – quasi in einer Black Box – abbilden zu können. Um isolierte Datensilos zu vermeiden, muss die Infrastruktur möglichst umfassende Informationen zum Kontext des Inhalts sowie zur Art der Teilkomponenten verstehen. Dafür ist eine formale semantische Beschreibung von Inhaltstypen in Form von Content Models erforderlich.

Objektmodell

Ein erstes abstraktes Objektmodell muss die Anforderungen an Versionierung, persistenter Identifikation, Relationen und Annotationen sowie Autorisierung erfüllen. Im Zuge der Implementierung verschiedener Anwendungen wird das Objektmodell verfeinert, um die disziplin-, institutions- und ressourcenspezifischen Charakteristika zu unterstützen.

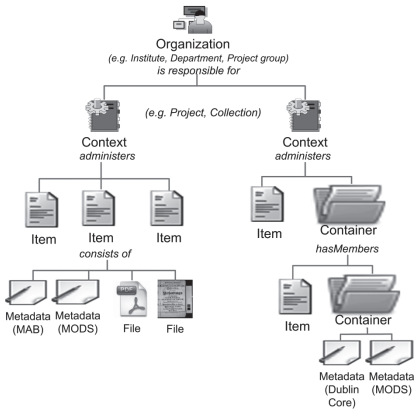

Abb.1 eSciDoc Objektmodell

Digital Preservation

Die längerfristige Speicherung binärer Datenströme sowie die technischen Anforderungen an Sicherheit, Emulation und Migration der Datenbestände sind von den zuständigen Einrichtungen bereits adressiert. Neben der Erhaltung der reinen Daten sind Informationen zur Identifikation und Semantik von Beständen, deren Kontext und Lebenszyklus für eine spätere Verwendung erforderlich.

Unabhängig von den organisatorischen Aspekten der Ablage von Forschungsdaten unterstützt die eSciDoc Infrastruktur die technische Langlebigkeit der Datenhaltung, da für sämtliche Daten die entsprechenden Informationen zu Kontext, Eigentümern, Historie bzw. Lebenszyklus in offenen und standardisierten Formaten abgelegt werden.

Vorgehen im eSciDoc Projekt

Basierend auf der skizzierten Ausgangslage, den daraus resultierenden technischen Anforderungen sowie den besonderen Zielsetzungen innerhalb der MPG ergaben sich für das eSciDoc-Projekt, neben den technischen Aspekten, bestimmte Prämissen, die die Projektkultur, Arbeitsweisen und Projektergebnisse beeinflussten.

Forschungs- und Ressourcenzentriert

Sämtliche Entwicklungen gehen auf konkrete, wissenschaftlich basierte Szenarien einzelner Fachdisziplinen zurück. Die Entwicklung der eSciDoc-Anwendungen erfordert eine zeitnahe und intensive Auseinandersetzung mit den disziplinspezifischen Artefakten, Arbeits- und Kommunikationsprozessen, weshalb Anwendungen im engen und ständigen Austausch mit den Nutzenden konzipiert werden.

Zu den grundlegenden Fragestellungen in der Phase der Konzeption und des Designs einer Anwendung zählen Typologie und Charakteristika der Artefakte sowie notwendiger und sinnvoller Granularität der Informationseinheiten bzw. ihrer Aggregationen. Im Zusammenhang des Lebenszyklus von bestimmten Ressourcen spielt die Definition von offenen und geschützten Artefakten eine wichtige Rolle. Das Verständnis für gewohnte Kommunikationskanäle und -partner ermöglicht die Definition von Anforderungen an Auffindbarkeit und Interoperabilität, im Besonderen im Hinblick auf potentielle interdisziplinäre Nachnutzung in bekannten und neuen Kontexten (re-use und re-purposing). Ein Kernstück des wissenschaftlichen Arbeitsablaufes sind angewandte Methoden und Werkzeuge, um Artefakte zu generieren, zu erschließen, zu bearbeiten und zu publizieren. Dabei stellt sich oft die Herausforderung, bereits bestehende disziplinspezifische, oft proprietäre Werkzeuge in offene Systeme zu integrieren.

Interaktion von Anwendung und Infrastruktur

Die Entwicklung von forschungszentrierten Anwendungen steht in ständigem Austausch mit der parallelen Entwicklung der Infrastruktur. Beide Ebenen bedingen einander und dienen dem gegenseitigen Ausbau: Auf der fachlichen Ebene werden Arbeitsprozesse, Artefakte und Werkzeuge identifiziert sowie auf mögliche generische und disziplinspezifische Komponenten hin untersucht. Komplementär werden auf der technischen Ebene generische Services und disziplinspezifische Solutions entwickelt. Wie bei jedem Infrastrukturprojekt ist die Herausforderung gegeben, an „Rohbau“ und „Innenausbau“ gleichzeitig zu arbeiten und die resultierenden Implikationen rechtzeitig und zeitnah zu erkennen.

Derzeit werden drei eSciDoc-basierte Anwendungen entwickelt, die jeweils verschiedene Szenarien und Artefakte unterstützen. Die Anwendung PubMan adressiert das Management und die Dissemination von Publikationsdaten. Komplementär dazu werden zwei Lösungen für das Szenario einer „Scholarly Workbench“ entwickelt: FACES beinhaltet den nachhaltigen Umgang und die Arbeit mit Bilddaten, VIRR adressiert die nachhaltige Ansichts- und Editionsumgebung für digitalisierte Textressourcen.[6] Alle drei Anwendungen haben eine solutionspezifische Logik, greifen jedoch auf generische Services zurück (z. B. import service).

Open Source & Community Building

Die Entwicklungen des eSciDoc-Projekts stehen sämtlich als Open Source frei zur Verfügung. Als Lizenz wurde die CDDL[7] gewählt. Nach vier Jahren intensiver Community-Arbeit innerhalb der MPG, der Identifikation und Konsolidierung von Early Adoptern, Piloten, Partnern und Mentoren auf allen Ebenen der MPG, wendet sich das eSciDoc-Projekt nun dezidiert der internationalen Community zu, um die Ergebnisse für neue Einsatzmöglichkeiten und Szenarien zur Verfügung zu stellen sowie die Nachhaltigkeit der Entwicklung zu gewährleisten.[8] Das Projekt muss sich seitdem stärker den Aspekten stabiler Interfaces, einem transparenten Entwicklungsvorgehen, den Fragestellungen zur Einbindung externer Entwickler und der stetigen Verbesserung der allgemeinen Kommunikation stellen. Die zahlreichen Anfragen von nationalen und internationalen Organisationen an Einsatz- und Entwicklungsmöglichkeiten von eSciDoc[9] erfordern ein entsprechendes Management der Erwartungshaltungen bzw. umfassender und vorausschauender Entwicklungsplanung. Als Ziel für die eSciDoc Days in 2009 steht deshalb eine verbesserte Transparenz in der Planung und Organisation von speziellen Anwendungsgruppen. Die diversen Anwendungsinteressen der MPG und der externen Interessenten werden in gesonderten Arbeitsgruppen oder „Special Interest Groups“ (z. B. zur Bilddissemination) adressiert und gemeinsam bearbeitet.

6 Weitere Angaben zu den einzelnen Anwendungen finden sich im CoLab Wiki der MPDL http://colab.mpdl.mpg.de/mediawiki/ESciDoc_Solutions_summary

7 http://www.sun.com/cddl/. Als Lizenz mit beschränktem Copyleft sichert sie zum einen die weitere freie Verfügbarkeit ab, schränkt jedoch nicht die Erstellung von eingeschränkt nutzbaren Lösungen ein, die z. B. Patentrechte berühren. Ein weiterer Vorteil der CDDL ist die Möglichkeit zur freien Benennung eines Gerichtsstandorts, ohne hierbei eine neue Lizenz erzeugen zu müssen, wie dies z. B. bei der MPL (http://www.mozilla.org/MPL/MPL-1.1.html) der Fall ist.

8 Als Startschuss zum Aufbau einer Open Source Community wurden im Juni 2008 die eSciDoc Days mit über 100 internationalen Teilnehmenden in Berlin abgehalten. Hierbei konnten durch die unterschiedlichen Ideen und Vorstellungen der Teilnehmenden zu einem Einsatz von eSciDoc neue Szenarien identifiziert werden, für welche die Infrastruktur eingesetzt werden kann.

9 Zur Zeit evaluieren in etwa 20 internationale Organisationen den Einsatz von eSciDoc.

Zusammenfassung und Rückblick

Der gewählte Ansatz für den Aufbau einer strategischen Infrastruktur hat sich bislang hinsichtlich gewählter Technologien, Transparenz in der Planung und enger Zusammenarbeit mit den Instituten als geeignet für die Anforderungen der MPG gezeigt. Das strategische Ziel, eine nachhaltige eScience-Umgebung aufzubauen, steht als Top-Down Entscheidung in einem potentiellen Konflikt mit den gewohnt Bottom-Up Ansätzen in wissenschaftlicher IT-Entwicklung. Umso dringender ist es, bewusste Balance zwischen schnellen, sichtbaren Mehrwertdiensten für einzelne Fachdisziplinen und den langfristigen, oft nicht deutlich wahrnehmbaren Investitionen in Nachhaltigkeit und Effizienz einer Infrastruktur zu halten. Die oft unterschiedlichen Anforderungen der Organisation und der Anwender erfordern ein sorgfältiges und diplomatisches Management der Erwartungen sowie transparente Projektplanung. Der gewählte Ansatz, frühzeitig interessierte und vor allem engagierte „Early Adopters“ zu finden, die als Fürsprecher der einzelnen Disziplinen auftreten, war dabei sehr hilfreich. Die frühe und fokussierte Einbindung der Fachdisziplinen sowie anderer Stakeholder ermöglicht parallel den Aufbau einer „Knowledge Infrastruktur“, die eine Begleiterscheinung einer eSciene-Organisation ist: Kompetenzen und Expertise über Artefakte, wie z. B. Bilddaten, sind über die Disziplinen als auch über traditionelle Expertiseträger (wie Wissenschaftler, IT und Bibliotheken) verteilt und es erfordert entsprechende Netzwerkarbeit, um implizites und explizites Wissen zusammenzubringen.

Die Anforderungen des Projekts erfordern entsprechend flexible interne Prozesse zum Projektmanagement und Expertise-Aufbau: vom Team ist nicht nur ein „Abarbeiten“ von Anforderungen gefordert, sondern vor allem ein „Erarbeiten“ von Expertise in fachlichen und technischen eScience-Fragestellungen zur Ableitung einzelner Dienste. Die unterschiedlichen Ziele und Vorgehensweisen für notwendige analytische und evaluative Konzeptions- und Architekturfragen und „harte“ Terminierung fertiger, reifer Produkte erfordern ein hohes Maß an intrinsischer Motivation und konstruktiver Mitarbeit des Teams. Die Projektplanung selbst muss von beiden Partnern flexibel genug gestaltet werden, um das Potential von unerwarteter, aber sinnvoller Nutzung der Infrastruktur und ihrer Komponenten auszuschöpfen: „be prepared for the unexpected“.

Nur kurz soll an dieser Stelle auch auf die Probleme der Komplexität größerer Software-Entwicklungsvorhaben hingewiesen werden. Im Projekt waren zeitweise über 20 Personen mit der reinen Softwareentwicklung beschäftigt, was eine feingranulare Planung der einzelnen Schritte, eine systemunterstützte Prozesssteuerung sowie strikte Vereinbarungen zu den Softwareentwicklungsumgebungen erfordert. Gleichzeitig hat der Aufbau dieser formalen Strukturen einen robusten softwaretechnischen Grundstein für die nun immer verteiltere Entwicklung gelegt.

Unabhängig von der strategischen Zielsetzung und den konkreten Anforderungen einer Organisation stellt sich für Entscheidungsträger die Frage der Nachhaltigkeit größerer Infrastrukturprojekte. Ein Projekt wie eSciDoc bietet durch den erheblichen Ressourceneinsatz in der Aufbauphase einen guten Ausgangspunkt für weiteren Ausbau und Entwicklungen, erfordert jedoch auch Beiträge vieler unterschiedlicher Institutionen im Sinne eines Open-Source-Community-Ansatzes, um die Einsatzmöglichkeiten und die Servicevielfalt zu steigern.

Die vielleicht deutlichste Herausforderung von Infrastrukturprojekten ist der implizite Bedarf an paralleler Organisationsentwicklung, die eigenen, oft nicht direkt beeinflussbaren Mechanismen unterliegt. IT-Systeme modellieren und greifen somit in bestehende Abläufe ein, das gilt umso mehr für übergreifende Infrastrukturprojekte, die sämtliche Bereiche einer Organisation berühren. Sowohl technische als auch organisatorische und soziale Strukturen müssen neu gedacht und konzipiert werden, was sich am deutlichsten bei Themen wie Digital Curation, persistenter Identifikation, Authentication/Authorisation oder Implementierung von Open-Access-Workflows zeigt. Hier stellt sich nicht nur die Frage nach Technologien, sondern auch die Frage nach nachhaltigen und kosteneffizienten Lösungen für das Zusammenspiel von zentralen und lokalen bzw. fachspezifischen Systemen, Expertisen und nicht zuletzt Personal- und Sachmitteln.