Open Source Software in der Archivierung

Archivierung

Christoph Jeggle

P R O J E C T C O N S U L T

Unternehmensberatung Dr. Ulrich Kampffmeyer GmbH

Hamburg 2005

Open Source Software in der Archivierung

Von Christoph Jeggle

Christoph Jeggle ist Seniorberater bei der PROJECT CONSULT Unternehmensberatung GmbH, Hamburg.

Was versteht man eigentlich unter OpenSource?

Was vor vielen Jahren als Zusammenarbeit von begeisterten Softwareentwicklern im universitären Unix Umfeld begann, ohne dass es kommerzielle Relevanz hatte, ist inzwischen zu einem wichtigen Faktor im Softwaremarkt geworden. Die so genannte Open Source Software.

Open Source Software ist freie Software, aber nicht Freeware. Freeware ist proprietäre Software, die kostenlos abgegeben und kopiert werden darf, ohne dass der Quellcode preisgegeben wird und ohne dass Veränderungen an der Software möglich sind. Freie Software soll das aber genau ermöglichen. „Freie Software ist Software, die mit der Erlaubnis für jeden verbunden ist, sie zu benutzen, zu kopieren und zu verbreiten, entweder unverändert oder verändert, entweder gratis oder gegen ein Entgelt.“(Free Software Foundation (2001): Kategorien freier und unfreier Software, online http://www.gnu.org). Open Source oder Freie Software, kann also durchaus kommerziell sein. Es ist zunächst nur die Aussage, dass die Software „quelloffen“ ist. Die Programmierung der Software wird offen gelegt und die Software kann von Programmierern verändert und neu kompiliert und gelinkt werden und so zu einem neuen, ablauffähigen Programm zusammengestellt werden. Insbesondere die Linux Plattform, ihrerseits ein Open Source Betriebssystem, ermöglicht diese Vorgehensweise. Die notwendigen Werkzeuge wie Compiler werden dazu mitliefert. Aber auch auf anderen Plattformen, wie z.B. Windows, ist Open Source Software inzwischen häufig anzutreffen. Wer sich mit dem Quellcode aber nicht auseinandersetzen möchte, erhält inzwischen bei vielen Open Source Produkten eine komfortable Installation, die sich nicht von proprietären Produkten unterscheidet. Z.B. die Installationsroutinen bekannter Linux Distributionen wie Suse, Red Hat besitzen eine grafische Benutzeroberfläche und führen einigermaßen verständlich durch die Installation.

Aber was ist nun das Besondere an Open Source Software, wenn sie sich inzwischen auch bei der Installation, der Konfiguration und der Funktionalität nicht mehr von proprietärer Software unterscheidet?

Durch die Veröffentlichung des Quellcodes laden die Herausgeber der Software dazu ein, die Software zu verändern und den eigenen Bedürfnissen anzupassen und diese Ergebnisse wieder als Open Source der Allgemeinheit bereitzustellen.

Dagegen wird der Quellcode proprietärer Software gehütet wie ein heiliger Gral, um keinem Mitbewerber zu ermöglichen, diese Software ohne großen Aufwand nachzubauen. Allerdings bedeutet dieser Schutz, dass die Software nur im Rahmen der vom Hersteller vorgesehen Funktionalität an die individuellen Bedürfnisse der

Anwender angepasst werden kann. In der Regel sind das Konfigurationsmöglichkeiten und Schnittstellen zu anderen Programmen. Die Kernfunktionalität der Software selbst kann aber nicht verändert werden.

Anders bei der Open Source Software. Hier ist es möglich, durch den Zugriff auf den Quellcode das Programm selbst anzupassen. Das erfordert natürlich Programmierkenntnisse. Aber auch für Anwender ohne diese Kenntnisse bietet Open Source den Vorteil, dass sich die Produkte durch die Mitwirkung zahlreicher Anwender sehr dynamisch verändern und immer mehr Funktionalität annehmen. Zumindest gilt das für große Projekte wie z.B. die Apache Software Foundation (http://www.apache.org). Selbst wenn eine Veränderung der Software nicht geplant ist, können Fachleute dem Quellcode entnehmen, wie die Software funktioniert und welche möglichen Sicherheitsrisiken sie besitzt. Solche Sicherheitslücken können dann geschlossen werden. Oft ist es aber gar nicht notwendig, das selbst zu tun, weil die gesamte Entwicklergruppe, die sich dem Projekt verbunden fühlt, sich auch für mögliche Probleme in der Software verantwortlich fühlt und sie zu lösen versucht.

Wie oben bereits erwähnt, ist Open Source zwar keine Lizenzbezeichnung, aber dennoch gibt es im Open Source Bereich Lizenzmodelle, die die Ziele von Open Source sicherstellen sollen. Allerdings unterscheiden sich diese Lizenzen erheblich. Ein Überblick über die Lizenzen bietet die Open Source Initiative unter http://www.opensource.org/licenses/ . Die Open Source Initiative prüft die Lizenzen auch hinsichtlich ihrer Verträglichkeit mit dem Open Source Gedanken.

Besonders streng ist die GNU General Public License, die sicher stellen will, dass der Zugang zum Quellcode offen ist, dass die Software nur unter denselben Lizenzbedingungen kopiert und weitergegeben werden kann und dass das Programm verändert werden kann. Auch die veränderte Software darf nur unter dieser Lizenzbestimmung weitergegeben werden.

Durch die Bedingung, dass das Programm unter den gleichen Lizenzbedingungen weitergegeben werden muss, entspricht diese Lizenz genau den Vorstellungen der Free Software Foundation.

Um die kommerziellen Bedingungen für Open Source Software zu verbessern, hat die Ende der neunziger Jahre gegründete Open Source Initiative diese Bedingung gelockert und dahingehend geändert, dass die Software unter den gleichen Lizenzbedingungen weitergegeben werden sollte, aber nicht muss. Damit ist es möglich, Open Source Software zu verändern und unter anderen Lizenzbedingungen weiterzugeben bzw. zu verkaufen. Ausdrücklich erlaubt ist bei der Open Source Initiative auch, andere Software, die zusammen mit der Open Source Software geliefert wird, unter ganz andere Lizenzbedingungen zu stellen. Damit ist es möglich, Open Source und proprietäre Software in komplexen Produkten zu mischen. So verwendet Oracle als Web Server die Open Source Software Apache, ohne die eigenen Softwareteile als Open Source zu veröffentlichen.

An dieser Stelle soll Open Source Software speziell unter dem Gesichtspunkt der elektronischen Archivierung betrachtet werden. Vier Open Source Produkte, die als gemeinsames Merkmal den OAI Standard unterstützen, grenzen sich aus der Vielfalt der WCM-Open-Source-Produkte ab:

- Fedora
- DSpace
- CDSware
- EPrints

OAI heißt Open Archive Initiative und besteht aus einer Gruppe von Institutionen, hauptsächlich Bibliotheken und Archiven, die es sich zur Aufgabe gesetzt haben, einen Standard zu entwickeln, um den Austausch von Metadaten zwischen Bibliotheken bzw. Archiven zu erleichtern. Dabei ist der Standard Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) entwickelt worden. Dieser Standard ermöglicht es, mit einer einzigen Schnittstelle auf unterschiedliche Archive zugreifen zu können, um die Metadaten holen zu können. Diese Schnittstelle bietet keine Abfragemöglichkeit, sondern nur die Möglichkeit, Metadaten aus unterschiedlichen Archiven an einer einzigen Stelle zusammenzufassen, um sie dort verfügbar zu machen. Der Zugriff auf den hinter den Metadaten liegenden Inhalt (Dateien, Dokumente) ist in der Schnittstelle nicht vorgesehen. Technisch basiert der Standard auf HTTP und XML, da davon ausgegangen wird, dass die Archive im Internet verfügbar sind.

Bei diesem Standard und auch bei den oben bereits aufgelisteten Open Source Produkten liegt der Schwerpunkt nicht in der Archivierung im Sinne einer Langzeitspeicherung von Daten, sondern in der einfachen und einheitlichen Bereitstellung der Metadaten. Alle Produkte sind aber in der Lage, mit der entsprechenden Hardware auch eine Langzeitarchivierung zu ermöglichen. Eine Verwaltung der Speichersysteme gehört aber zu keinem der Produkte, kann aber durch die dokumentierte Speicherschnittstelle hinzugefügt werden.

Fedora

Fedora (http://www.fedora.info) zuletzt behandelt im Newsletter 20040817, darf nicht verwechselt werden mit der Red Hat Linux Distribution gleichen Namens. Beide haben nichts miteinander zu tun. Fedora heißt Flexible Extensible Digital Object and Repository Architecture und wird entwickelt von der Cornell University und der University of Virginia mit der finanziellen Unterstützung der Andrew W. Mellon Foundation, die die Weiterentwicklung mindestens bis 2007 garantiert. Einer der Chefdesiger, Carl Lagoze, ist übrigens zu Zeit im Executive Commitee der OAI.

Fedora liegt inzwischen in der Version 2.0 vor. Die erste Version wurde im Mai 2003 veröffentlicht.

Fedora kann als binäre Distribution heruntergeladen und installiert werden. Es kann aber auch nur der Quellcode herunter geladen und anschließend kompiliert werden. Beide Installationsarten sind gut beschrieben und einfach durchzuführen.

Da es sich um eine Java Implementierung handelt, ist Fedora für Unix, Linux und Windows verfügbar. Vorausgesetzt wird Sun's Java Software Development Kit, v1.4 oder höher. Fedora verwendet einen in der binären Distribution mitinstallierten Apache Tomcat 5 als Applikationsserver. Als Datenbanken können MySQL, v3.23.x, MySQL 4.x, oder Oracle 9i verwendet werden. Ist keine dieser Datenbanken vorhanden, kann auch die Open Source Datenbank McKoi mitinstalliert werden.

Fedora bietet im Wesentlichen folgende Funktionalität:

- Flexibles Modell für die digitalen Objekte.

  Beliebige digitale Objekte (Textdateien, Bilder, Videos, Webseiten,…) können im System verwaltet werden. Genauso gut ist das System in der Lage, nur Referenzen auf die Objekte zu verwalten, ohne dass die Objekte selbst in das System geladen werden müssen. Dabei kann unterschieden werden, ob sich Fedora um das Laden der referenzierten Objekte kümmern soll, wodurch der Link der aufrufenden Anwendung verborgen bleibt, oder ob die Anwendung nur den Link erhält und selbst für das Laden des Objektes verantwortlich ist.

  Unter dem Namen Disseminator wird eine besondere Funktionalität eingeführt, mit der Objekte mit Webservices verbunden werden können. Diese können unmittelbar zusammen mit dem Objekt aufgerufen werden. Ein übliches Anwendungsfeld ist das Erstellen von Renditions „on the fly“. Damit kann abhängig vom Aufruf des Objektes die Darstellung der Inhalte variiert werden.

- Versionierung

  Der Inhalt von digitalen Objekten unterliegt einer Versionierung. Der Aufruf einzelner Versionen ist möglich. Zusätzlich wird ein Audit Trail über die Art der Veränderungen geführt.

- XML Im- und Export

  Für den Im- und Export von Objekten kann zu Zeit zwischen zwei XML Formaten gewählt werden: Fedora Object XML (FOXML) und Metadata Encoding and Transmission Standards (METS). Zukünftig sollen die Digital Item Declaration Language MPEG2/DIDL und METS 1.3 unterstützt werden.

- Objektbeziehungen

  Objekte können untereinander in Beziehung gesetzt werden. Diese Beziehung wird als Triple in der Form von Subjekt, Prädikat und Objekt beschrieben. Damit lassen sich beliebige Beziehungen beschreiben. Beispiel: Objekt A (=Subjekt) ist ein Kapitel (=Prädikat) von Objekt B(=Objekt). Ermöglicht wird diese sehr flexible, in Graphen darstellbare Indizierung durch die Verwendung der Metadatenbank Kowari, ein Open Source Produkt von Tucana Technologies. Die Datenbank orientiert sich an den Vorgaben des W3C Resource Description Framework (RDF) Standards.

- Zugriffskontrolle

Diese Funktionalität ist zurzeit nur schwach ausgebildet und beruht auf einer Authentifizierung durch die IP Adresse. In der nächsten Version ist ein komplexeres System für die Zugriffskontrolle vorgesehen.

- Einfache Suche

  Ein einfacher, in Tabellen einer relationalen Datenbank gespeicherter Index verwaltet interne Objektmetadaten, die Fedora zur Verwaltung benötigt, und einen Standardindex, der auf dem Standard der Dublin Core Metadata Initiative (http://www.dublincore.org) basiert, die es sich zur Aufgabe gemacht hat, die gegenseitige Verarbeitbarkeit von Metadaten Modellen durch größtmögliche Standardisierung zu ermöglichen.

- RDF basierter Ressource Index

  Die bereits oben erwähnte Kowari Datenbank kann nicht nur für die Objekt-zu-Objekt Beziehungen eingesetzt werden, sondern auch für erweiterte Metadaten zu einzelnen Objekten.

Herausragende Features in dieser Liste sind die Verknüpfung von Webservices mit Objekten und die Verwendung der Kowari Datenbank für die Darstellung von Relationen zwischen den Objekten. Der Ansatz mit Web-Services erleichtert die Integration in SOA Service Oriented Architecture Lösungen.

Als Schnittstellen werden zwei SOAP basierte Webservices bereitgestellt, die Management und die Access API. In eingeschränkter Weise werden diese auch als HTTP-Service basierend auf REST (Representional State Transfer) bereitgestellt. Zusätzlich werden zwei Suchschnittstellen angeboten, eine für die einfache Suche in der relationalen Datenbank und eine für die Suche in der Kowari Datenbank.

Die Verwendung der oben genannten Schnittstellen in eigenen Anwendungen ist neben einigen Beispielanwendungen und einer Administrationsapplikation die einzige Möglichkeit, Fedora zu verwenden. Fedora ist also keine Software, die direkt nach der Installation bereits eingesetzt werden kann. Sie setzt immer einen Aufwand für die Integration in bereits bestehende oder neu zu erstellende Anwendungen voraus.

Fedora ist lizensiert mit der Mozilla Public License Version 1.1. Diese Lizenz erlaubt es, Fedora mit anderen Applikationen, die unter einer anderen Lizenz stehen, zu kombinieren.

DSpace

Die Open Source Software DSpace (http://www.dspace.org), zuletzt behandelt im Newsletter 20040817 ist eine gemeinsame Entwicklung des Massachusetts Institute of Technologies (MIT) und Hewlett-Packard (HP). Das Ziel dieser Software ist die Erfassung, Speicherung, Indizierung, Aufbewahrung und Weitergabe von Forschungsmaterial und anderen Dokumenten im digitalen Format.

Die erste Version ist im November 2002 veröffentlicht worden. Von vornherein ist das Projekt als Open Source angegangen worden, um die Erfahrungen mit anderen Forschungsinstitutionen teilen zu können und eine gemeinsame Weiterentwicklung zu ermöglichen. Aus der anfänglichen Gruppe von 7 Forschungsinstitutionen ist inzwischen eine Gemeinschaft von etwa 125 Institutionen geworden, die den Einsatz

von DSpace zur Veröffentlichung von Forschungsergebnissen und Dokumenten prüfen. 20 Universitäten setzen DSpace bereits produktiv ein. Die Weiterentwicklung von DSpace wird jetzt nicht mehr nur vom MIT und HP betrieben, sondern zunehmend von der wachsenden Gruppe der Anwender. Durch die Veröffentlichung als Open Source ist eine aktive Teilnahme an der Weiterentwicklung ermöglicht worden. Das Projekt wird ebenso wie Fedora von der Andrew W. Mellon Foundation unterstützt.

DSpace liegt zur Zeit in der Version 1.2.2 vor. Es setzt ein Unix oder Linux Betriebssystem voraus. Eine Windows-Version steht offiziell nicht zur Verfügung, aber durch die Systemarchitektur ist eine Plattformunabhängigkeit gegeben. Daher kann DSpace auch unter Windows installiert werden. Wie Fedora basiert auch DSpace auf einem Apache Tomcat als Applikationsserver. Folglich wird auch Java mindestens in der Version 1.4 vorausgesetzt. Als Datenbank wird PostgreSQL eingesetzt. Aus den Konfigurationsdateien ist aber zu entnehmen, dass auch ein Einsatz von Oracle (in Zukunft) möglich sein soll. Die Installationsanweisungen beziehen sich aber ausschließlich auf PostgreSQL. Die Anweisungen sind ausführlich, setzen aber gewisse Kenntnisse bezüglich Tomcat und PostgreSQL voraus. Die Installation erfolgt immer manuell und nicht über eine automatische Setuproutine.

DSpace bietet folgende Funktionalität:

- Datenorganisation

  Die Art und Weise, wie DSpace die Daten organisiert, soll die Organisation der Institutionen nachbilden. An oberster Stelle stehen die Communities, die selbst wieder hierarchisch organisiert werden können. Jede Community bzw. SubCommunity hat Collections, die z.B. Themengebiete zusammenfassen. Collections können zu mehreren Communities gehören. Das eigentliche Archivelement ist das Item, das zu einer Collection gehört, aber in mehreren Collections referenziert werden kann. Zum Item gehören die Metadaten des Archivobjektes. Die Items selbst sind wieder organisiert in Bundles von Bitstreams. Durch die Bundles wird ermöglicht, dass Archivobjekte aus mehr als nur einer Datei bestehen können (z.B. Webseiten mit den dazugehörigen Bildern), denn die Bundles bestehen aus einem oder mehreren Bitstreams, den eigentlichen Daten. Das Item kann aber aus mehr als einem Bundle bestehen. Eine gebräuchliche Einteilung der Bundles ist: Original, Thumbnails, Text (für die Indizierung). Zu jedem Bitstream muss auch das Bitstream Format angegeben werden, das den MIME Typ und die Ebenen des Supports (unterstützt, bekannt, nicht unterstützt) beschreibt.

- Metadaten

  Der Standardumfang der Metadaten eines Items umfasst den Dublin Core, zuzüglich einiger beschreibender Informationen für die anderen Ebenen der Datenorganisation. Die Metadaten werden noch ergänzt um administrative Metadaten, die sich auf die Herkunft und Aufbewahrung der Daten beziehen und strukturelle Metadaten, die die Beziehung der Bitstreams untereinander beschreiben.

- Benutzerverwaltung

  Auf jeder Ebene der Datenorganisation können Rechte an Benutzer und Gruppen vergeben werden. Diese Rechte werden pro Ebene vergeben und nicht weiter auf die nächste darunter liegende Ebene weitergegeben. Ein anonymer Zugriff auf Elemente des Systems kann freigegeben werden. Die Authentifizierung kann entweder über eine Kombination, über Name und Passwort oder über eine X509 Zertifizierung durchgeführt werden. Die eingetragenen Benutzer können auch benachrichtigt werden, wenn in definierten Bereichen neue Dokumente eingestellt werden.

- Dokumente in das System einstellen

  Dieser Prozess kann sowohl über eine Stapelverarbeitung, als auch über eine Web-Benutzerschnittstelle für einzelne Dokumente durchgeführt werden. Es können bis zu drei Workflow-Schritte definiert werden, die den Freigabeprozess darstellen.

- Global Unique Identifier

  Der Zugriff auf die Daten geschieht über eine URL. Da diese einem Wandel unterworfen ist durch die Verschiebung von Webseiten und Ähnliches, macht es Sinn, ein System für die global eindeutige Identifizierung von Objekten zu haben. Dieses System wird von der Corporation for National Research Initiatives (CNRI) bereitgestellt. DSpace unterstützt dieses System.

- Suchen und Blättern

  Unterstützt wird die strukturierte Suche nach Metadaten und die Volltextsuche. Als Volltextengine wird die Open Source Software Lucene verwendet.

- Import und Export

  Neben der bereits erwähnten Stapelverarbeitung für das Einfügen von neuen Daten können Daten auch exportiert werden. Dabei wird das METS Format (siehe oben bei Fedora) verwendet.

- History

  Ereignisse werden auf die Objekte bezogen protokolliert. Diese Funktion befindet sich aber noch in einem relativ ungetesteten Zustand.

Bemerkenswert ist, dass DSpace im Gegensatz zu Fedora sofort ein gebrauchsfertiges System liefert, das zwar noch konfiguriert werden, aber nicht mehr unbedingt durch Programmierung erweitert werden muss. Ein wichtiges Feature des Systems ist auch die standardmäßige Unterstützung eines einfachen Freigabeprozesses.

Als Programmierschnittstellen werden Java Klassen auf drei verschiedenen Ebenen bereitgestellt, die aufeinander aufbauen: Storage Layer, Business Logic Layer, Application Layer.

DSpace ist lizensiert mit einer Open Source Lizenz, die an die BSD Lizenz angelehnt ist.

Wie die zuvor behandelten Produkte sind beide ebenfalls OAI PMH kompatibel. Das OAI PMH, Open Archive Initiative Protocol for Metadata Harvesting, ermöglicht, die Metadaten unterschiedlicher Archive in ein gemeinsames Verzeichnis holen zu können. OAI PMH ist keine Abfragesprache, sondern ein Protokoll zur Bildung archivübergreifender Verzeichnisse.

CDSware

CDSware (http://cdsware.cern.ch) bedeutet CERN Document Server Software und ist tatsächlich eine Entwicklung der Europäischen Organisation für nukleare Forschung. Gegründet worden ist die Organisation allerdings als "Conseil Européen pour la Recherche Nucléaire", daher die Abkürzung CERN. Genauer beschrieben wird die Tätigkeit des CERN durch die Bezeichnung als Europäischen Laboratorium für Teilchenphysik.

Bekannt geworden ist das CERN auch dadurch, dass Tim Berners-Lee und ein kleines Team am CERN das World Wide Web entwickelt haben.

CDSware ist am CERN für den eigenen Bedarf entwickelt worden und unter der GNU General Public Licence (GPL) als freie Software veröffentlicht worden. Es steht damit auch anderen zur Installation und Benutzung einschließlich Sourcecode zur Verfügung. Am CERN wird CDSware als Online-Katalog von Veröffentlichungen unterschiedlicher Art verwendet und umfasst über 650.000 Einträge mit 320.000 Volltext Dokumenten. Der CERN Document Server ist öffentlich und kann unter der Adresse http://cds.cern.ch erreicht werden.

Dort sind dann auch die wesentlichen Features dieses Dokumenten Servers zu erkennen. Er stellt zunächst einmal eine portalähnliche Benutzerschnittstelle im Webbrowser zur Verfügung, Die Ähnlichkeit zu einem Portal ergibt sich auch durch die Möglichkeit der Personalisierung und der Abspeicherung von ausgewählten Dokumenteinträgen in benutzerspezifischen „Dokumentkörben“. Dieses Portal dient auch als Eingabeplattform für eine leistungsstarke Suchmaschine, die den Katalog der verwalteten Dokumente durchsucht. Die Suchsyntax ist der von Google ähnlich. Sowohl die Suche nach einzelnen Feldern als auch die Volltextsuche ist möglich.

Über die Benutzerschnittstelle ist es ebenfalls möglich, mit der entsprechenden Authentifizierung und Autorisierung Dokumente und ihre Metadaten in das System einzustellen.

Technisch gesehen basiert CDSware auf einem UNIX oder Linux System mit einem Apache Webserver und einer MySQL Datenbank. PHP und Python Unterstützung für den Apache Webserver muss mitinstalliert werden. Insgesamt ist die Installation nicht ganz einfach, da sie nicht durch ein Installationsprogramm unterstützt wird, sondern manuell Schritt für Schritt durchgeführt werden muss. Dabei sind die Schritte aber ausreichend dokumentiert. Einige Komponenten müssen eine bestimmte Version haben (z.B. das Python Modul für MySQL), was nicht unbedingt die neueste Version bedeutet. Das kann bei neueren Linux Distributionen zu allerdings lösbaren Komplikationen durch die Installation von älteren Komponenten führen.

Einen Blick sollten wir im Zusammenhang mit CDSware noch auf das Format der Metadaten werfen. CDSware verwendet MARC 21. MARC steht dabei für Machine-

Readable Cataloging. MARC 21 ist 1999 aus unterschiedlichen MARC Standards hervorgegangen und beschreibt ein maschinenlesbares Austauschformat. Das Format gliedert den Datensatz in den Leader, der bei der Verwendung von MARC 21 im bibliografischen Bereich allgemeine Angaben zur Art des erfassten Dokumentes oder der erfassten Daten macht, das Directory, das einen Index der verwendeten Felder einschließlich Feldkennung, Feldlänge und Position im Datensatz, und die eigentlichen Felder. Alle Elemente sind durch im Standard festgelegte Codes gekennzeichnet. Diese Codes können auch bei der Suche in CDSware eingesetzt werden, um die Suche näher einzugrenzen.

CDSware stellt ein vollständiges System für die Verwaltung bibliografischer Informationen einschließlich der dazugehörigen Dokumente dar. Es kann durch Module erweitert werden, wie es beim CERN Dokumenten Server zu sehen ist. Dort ist ein Konvertierungsmodul eingebunden, um Datenformate ändern zu können.

EPrints

EPrints (http://www.eprints.org) ist eine Entwicklung eines kleinen Teams an der Universität von Southampton. Die Codierung selbst wird im Wesentlichen von einem Entwickler geleistet. Das Projekt ist Teil des Open Citation Project, einem DLI2 International Digital Libraries Project unterstützt vom Joint Information Systems Committee (JISC) in Großbritannien. Zurzeit wird die Software in 161 Archiven eingesetzt.

Die Idee hinter diesem Projekt ist es, Forschern die Veröffentlichung ihrer Dokumente so einfach wie möglich zu machen. Dieses Konzept wird Self-Archiving genannt und meint im Grunde die Möglichkeit, über ein Web-Interface Dokumente selbst in das Archiv einstellen zu können.

Entsprechende Features werden bereitgestellt.

- Es ist möglich, unterschiedliche Dokumentenformate im System zu speichern. Ein Dokument kann gleichzeitig in unterschiedlichen Formaten abgelegt werden.
- Die Metadatenstruktur ist sehr flexibel. Es wird ein Pool von Metadatenattributen definiert, die dann einem oder mehreren Dokumententypen zugewiesen werden. Gleichzeitig ist es möglich eine Hierarchie von Themen (subjects) aufzubauen, unter denen Dokumente gefunden werden können. Für jedes Attribut kann entschieden werden, ob es verpflichtend ist und ob es für den OAI PMH Zugriff sichtbar ist.
- Das Einstellen von Dokumenten geschieht über das Webinterface. Es ist auch möglich, Dokumente als ZIP-Archiv im Bündel zu übergeben. Eine weitere Option ist die Übergabe von Dokumenten als Link. Über diesen Link holt EPrints dann das Dokument in das Archiv.
- Routinen, die die Datenintegrität sicherstellen, sind als Vorgabe bereits vorhanden, können aber angepasst und erweitert werden.
- Das Einstellen von Dokumenten kann durch einen Freigabeprozess erweitert werden

- Benachrichtigungen über neue Dokumente aus zuvor definierten Bereichen können per E-Mail zugestellt werden (subscription).

Ein Demonstrationsarchiv unter http://demoprints.eprints.org bietet die Möglichkeit, das System zu evaluieren.

Technisch basiert EPrints auf UNIX/Linux Systemen mit Apache als Webserver und MySQL als Datenbank. Die Programmierung erfolgt in Perl. Die Installation ist relativ komfortabel, da sie umfassend durch Skripte unterstützt wird.

EPrints wird veröffentlicht als freie Software und ist Teil des GNU Projektes (http://www.gnu.org).

Wie bereits CDSware und DSpace stellt auch EPrints ein System dar, das nach der Installation und Konfiguration sofort verwendet werden kann. Der Schwerpunkt liegt hier in der möglichst komfortablen Möglichkeit, Dokumente in das System einzustellen.

Zusammenfassung

Alle vier Produkte, die unter dem Begriff Archiv als Open Source Software veröffentlicht werden, kommen aus dem Bereich der Forschung und Lehre. Ihr Ziel ist es, eine Plattform zu bieten, um Dokumente beliebiger Art ablegen und sie dabei sinnvoll mit Metadaten versehen zu können. Die Flexibilität der Metadatenstruktur ist unterschiedlich. Während DSpace weitgehend nur auf dem Dublin Core beruht, können EPrints und CDSware erweitert werden. Besondere Stärken hinsichtlich der Metadaten zeigt Fedora, das zwar in seinem Grundindex auch auf dem Dublin Core beruht, aber über die RDF konforme Kowari Datenbank gerade im Bereich der Verknüpfung von Objekten sehr flexibel ist.

Fedora hat eine Sonderstellung in der Reihe der vorgestellten Archive, da die Software eine Webservice Schnittstelle zur Konvertierung und Bearbeitung der Inhalte beim Aufruf bietet. Allerdings ist Fedora auch das einzige Produkt in dieser Reihe, das eigentlich nur eine Infrastruktur und keine gebrauchsfertige Applikation bietet.

Allen Produkten ist gemeinsam, dass der Begriff Archiv hier nicht im Sinne von revisionssicherer Archivierung verwendet wird. Die Speicherung der Daten geschieht standardmäßig in Verzeichnissen auf lokalen oder im Netz verfügbaren Laufwerken. Eine Unterstützung von optischen Medien oder gar Jukeboxen ist standardmäßig nicht vorgesehen. Technisch wird es aber bei allen Produkten Möglichkeiten geben, das durch proprietäre Produkte aus dem Storagebereich zu ergänzen.

Alle vier Produkte stellen brauchbare Open Source Alternativen bei Aufgabenstellungen aus dem Bereich der Dokumentenablage und –veröffentlichung dar. Bei allen Produkten, auch denen, die relativ gebrauchsfertig installiert werden, darf der Aufwand zur Anpassung und Anbindung an bestehende Systeme nicht unterschätzt werden. Bei den Anpassungsarbeiten ist sehr sorgfältig vorzugehen, damit mit dem nächsten Update diese Arbeiten nicht hinfällig geworden sind.

Open Source Software in der Archivierung

Anschrift des Autors

PROJECT CONSULT GmbH, Büro Hamburg
Breitenfelder Str. 17
D-20251 Hamburg
Tel.: 040 / 460 762 20
Fax: 040 / 460 762 29
E-Mail: Presse@PROJECT-CONSULT.com
Web: www.PROJECT-CONSULT.com