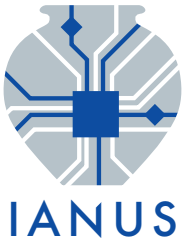
IANUS — Forschungsdatenzentrum
Archäologie & Altertumswissenschaften



c/o Deutsches
Archäologisches Institut
Podbielskiallee 69-71
D-14195 Berlin

Tel.: +49-(0)30-187711-359
ianus-fdz@dainst.de
www.ianus-fdz.de

TESTBED

„Softwaresuite des DA-NRW“

ABSCHLUSSBERICHT (HENDRIK SCHMEER)

25. Juni 2014

EINLEITUNG

Im Rahmen eines Testbeds wurde von Oktober 2013 bis April 2014 durch den Autor die Softwaresuite des Digitalen Archives NRW (DA-NRW) hinsichtlich der Nachnutzbarkeit für IANUS evaluiert. Zu diesem Zweck wurde das Paket auf einem Virtuellen Server im Rechenzentrum der Universität Köln installiert und konfiguriert. Zuletzt fanden Tests mit Datensammlungen statt, die IANUS von verschiedenen Datengebern für interne Zwecke erhalten hat, um einen künftigen Bedarf für Anpassungen und Eigenentwicklungen besser abschätzen zu können.

Koordination
Deutsches
Archäologisches
Institut

Förderung
Deutsche
Forschungsgemeinschaft
DFG

Die DA-NRW Softwaresuite wurde aus mehreren quelloffenen Softwaresystemen und Eigenentwicklungen für die Bedürfnisse des Digitalen Archivs NRW[1] entwickelt. Da das DA-NRW primär als spartenübergreifende Infrastruktur für institutionelle Einrichtungen und Dateneigentümer des Landes Nordrhein-Westfalen (insbesondere Bibliotheken, Archive, Museen, etc.) gedacht ist, werden die Anforderungen an wissenschaftliche Daten oder individuelle Datengeber nur partiell unterstützt. Im Folgenden werden die einzelnen Komponenten der Softwaresuite erläutert.

CONTRACTORS & ACCESS MANAGEMENT

Der LZA Bereich des DA-NRW verfügt über kein besonders differenziertes System zur Steuerung des Zugangs; es unterscheidet lediglich zwischen verschiedenen Kontraktoren und der Öffentlichkeit. Ein Kontraktor bezeichnet in der Regel eine ganze Institution, insofern eher eine Benutzergruppe, die unter dem Namen der Institution Pakete einliefert und als „Eigentümer" besondere Rechte über die eingelieferten Daten erhält. Alle anderen Kontraktoren werden als Öffentlichkeit angesehen.

FAZIT: Dieses Merkmal erscheint für IANUS ungeeignet, da es zu wenig granular ist und zu sehr auf institutionelle und nicht auf individuelle Datengeber ausgelegt ist. Der Zugriff auf die Präsentationsebene (Fedora), die auch die Schnittstelle zur Suche zur Verfügung stellt, kann jedoch anders gesteuert werden.

IRODS

Das OpenSource Software Paket stellt die nötigen Funktionalitäten zur Verfügung, um Daten über mehrere, entfernt stehende Knoten (i.d.R. Archiv-Server in Rechenzentren) zu replizieren und verfügbar zu halten. Die Verwaltung dieses Systems geschieht zentral über eine Datenbank auf einem der Knoten (iCAT Server). Die Speicherorte der Daten werden als logische Pfade unabhängig von der tatsächlichen physischen Repräsentation abgebildet, die realen Pfade bestimmt das System beim Zugriff intern.

Vom Prinzip her wird eine vollständige Spiegelung der Knoten untereinander angestrebt. Jedoch ist dies insbesondere beim Einsatz langsamer Bandlaufwerke als Archivspeicher nicht immer der Fall. Aus diesem Grund werden die Knoten beim Abruf nicht individuell wie Spiegelserver angesprochen, sondern iRODS entscheidet intern, woher die Daten zu beziehen sind.

[1] http://da-nrw.hki.uni-koeln.de/projects/danrwpublic - Eine ausführliche Beschreibung bietet: M. Thaller (Hrsg.), Das Digitale Archiv NRW in der Praxis. Eine Softwarelösung zur digitalen Langzeitarchivierung (Hamburg 2013).

Insofern stellt dieses System, das auch den Zugriff verwaltet, gegenüber einer bloßen Spiegelung von Servern einen Vorteil dar, sobald das Archiv stetig um große Datenmengen über Annahmestellen unterschiedlicher Knoten erweitert wird, deren Replikation einige Zeit in Anspruch nehmen kann. In dem Fall bestünde für den Benutzer große Unsicherheit darüber, an welchem Knoten er auf bestimmte Daten zugreifen könnte.

Ist diese Software einmal installiert und richtig konfiguriert, ist der weitere Wartungsaufwand gering, Probleme sind während des bisherigen Testbetriebs oder innerhalb des DA-NRW nicht aufgefallen. Die Erweiterung des Grids um weitere Ressourcen zur Datenreplikation gestaltet sich dann sehr einfach.

FAZIT: Da momentan für IANUS nur eine Archivierung und Bit-Stream-Preservation bei zwei verschiedenen Rechenzentren (Knoten) vorgesehen ist und der Datenzuwachs zunächst überschaubar sein wird, erscheint eine Spiegelung der Bandlaufwerke über klassische Methoden ausreichend und der Mehrwert von iRODS eher gering. Jedoch sind die Softwareschnittstellen für iRODS in der DA-NRW Softwaresuite bereits vorhanden, wohingegen andere Adapter erst noch programmiert werden müssten. Zudem erfüllt iRODS alle Voraussetzungen im Hinblick auf eine spätere Skalierbarkeit.

CONTENTBROKER

Der ContentBroker ist eine Entwicklung des Instituts für Historisch-kulturwissenschaftliche Informationsverarbeitung (Lehrstuhl Prof. M. Thaller) an der Universität zu Köln. Er erkennt an den Eingabeschnittstellen (*ingest area*) eingelieferte *Submission Information Packages (SIP)* und steuert maßgeblich die Umwandlung in *Archival Information Packages (AIP)* und *Presentation Information Packages (PIP)*. Hierbei werden unter anderem Dateiformate ggf. konvertiert, oder Daten für die Präsentation vorbereitet. Beide Vorgänge müssen über sogenannte *Conversion Strategies* in der Datenbank des ContentBroker konfiguriert werden. Außerdem werden während der Umwandlung zum AIP technische Metadaten ergänzt und METS Metadaten, so vorhanden, nach Dublin Core gemappt.

Dazu werden Formate zunächst über die Fileformat Registry PRONOM[2] identifiziert und über die hier vorgefundenen Identifier wird eine geeignete *Conversion Strategy* angesprochen, so vorhanden. Anderenfalls wird die eingelieferte Datei ohne weitere Konvertierung archiviert. Technisch wird etwa zur Konversion von Bildformaten die Programmbibliothek Imagemagick eingebunden.

Die meisten Transformationen lassen sich über *„CLI Conversion Strategies“* abdecken, über die jedes unter Linux lauffähige Kommandozeilenprogramm angesprochen werden kann. Auf diese Weise ist der ContentBroker sehr flexibel einsetzbar, ohne den Quellcode zu verändern.

[2] http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new

Eine Limitation sind an dieser Stelle jedoch Formate, die nicht auf einem Linux System behandelt werden können, da die nötige Software für dieses OS nicht zur Verfügung steht – wie beispielsweise viele proprietäre Formate wie von Microsoft oder Adobe. Für diesen Zweck betreibt das DA-NRW einen eigenen Server mit Microsoft Betriebssystem, auf dem dann die benötigte Transformation ausgeführt werden kann.

Eine weitere Einschränkung der *CLI Conversion Strategies* sind komplexere Aufgaben, bei denen beispielsweise noch zusätzliche Parameter aus den Eingabedaten ausgelesen werden müssen oder weitere Logik vor dem eigentlichen Programmaufruf abgearbeitet werden soll. Zu diesem Zweck können spezielle *Conversion Strategies* in Java programmiert werden, die dann über ihren Namen einzubinden sind. Innerhalb der DA-NRW Softwaresuite findet dies bei vielen Transformationen für das *Presentational Repository* Anwendung, wo zum Beispiel Bilder für die Präsentation kleiner dargestellt, mit einem Wasserzeichen versehen oder Videos auf eine einleitende Sequenz beschnitten werden.

Als letzten Schritt in der Arbeitsabfolge erledigt der ContentBroker auch das Zurückholen der eingelieferten Pakete aus dem Archiv, das sogenannte „Retrieval".

FAZIT: Im Kern ist der ContentBroker für IANUS geeignet, da er standardkonform und sicher Konvertierungen von Formaten vornimmt und die Veränderungen an Dateien protokolliert. Anpassungsaufwand besteht bezgl. der noch zu definierenden Migrationsstrategien, da sehr viele unterschiedliche Dateiformate in den Altertumswissenschaften zum Einsatz kommen.

Nachteilig ist, dass der ContentBroker in seiner jetzigen Form agnostisch gegenüber nicht-technischen Metadaten und Beschreibungen ist, also nicht überprüft, ob inhaltliche Mindestangaben vorhanden sind oder eine nach formalen Kriterien ausreichende Dokumentation vorliegt. Diese Aufgabe muss erfüllt werden, bevor Daten durch den ContenBroker prozessiert werden.

FEDORA REPOSITORY

Diese Software wird hier als *Presentation Repository* eingesetzt, um die archivierten Daten durchsuchen zu können. Fedora ist in der Lage, Metadaten des Formats Dublin Core zu verarbeiten und zu indizieren. Grundlage der Suche sind die beim Ingest eingelieferten Metadaten in den oben genannten Formaten, bzw. Mappings auf Dublin Core. Aus dem Grund der Beschränkung auf Dublin Core auch in der Suchfunktion wird in DA NRW zusätzlich noch *ElasticSearch* als weiterer Suchindex mit eingebunden.

Innerhalb von Fedora kann auf die zur Präsentation vorbereiteten Daten, bzw. vom ContentBroker erstellten und in Fedora eingelieferten PIPs zugegriffen werden. Die von Fedora bereitgestellte Datenstruktur ist nicht redundant und

auch nicht zur Langzeitsicherung geeignet. Vielmehr handelt es sich hier um ein eigenständiges System, das mit PIPs aus dem Archivsystem beliefert wird.

Das Einsehen der in Fedora bereitgestellten Daten ist nicht zu verwechseln mit dem „Retrieval“ eines archivierten Paketes. Die präsentierten Daten werden oftmals kleiner dargestellt oder nur ausschnittsweise. Zusätzlich können für verschiedene Benutzergruppen (interne, externe, die Öffentlichkeit) unterschiedliche PIPs erstellt werden, deren Detailtiefe angepasst werden kann.

FAZIT: Fedora ist ein äußerst komplexes und flexibles Framework, dessen Möglichkeiten und Grenzen innerhalb dieser Evaluation kaum ermittelt werden können. Zum Aufbau eines funktionierenden Systems sind sowohl zeitliche als auch personelle Anstrengungen erforderlich. Es scheint, als wäre auch eine sehr differenzierte Steuerung des individuellen Objektzugriffs bis auf die Ebene der Metadaten über Fedora und XACML (*eXtensible Access Control Markup Language*) möglich.

SIP BUILDER

Diese Software ist ein Werkzeug zur Erstellung von einlieferungsfähigen Paketen. Im Wesentlichen werden Dateien komprimiert, mit Prüfsummen versehen und in das *Bagit* Format gepackt. Der SIP-Builder ist dabei Dateiformat-agnostisch.

Bei der Erzeugung des SIP können Richtlinien zur Präsentation bzw. für PIPs festgelegt werden. Die Präsentationsrichtlinien unterscheiden zwischen einer Präsentation für interne Anwender (eigene Institution) und Externe (die Öffentlichkeit). Es lassen sich Einschränkungen zur zeitlichen Verfügbarkeit angeben, Größenbeschränkungen bei Bildern oder Ausschnitte bei Videos.

FAZIT: Das Manko am SIP-Builder im Bezug auf IANUS: Die Möglichkeiten zur Einschränkung der Präsentation von *Dissemination Information Packages* müssten weiter ausdifferenziert werden können, z.B. dass diese nur für einzelne Dateien oder die gesamte Datensammlung gelten.

DA-WEB

DA-Web ist ein unter Tomcat laufendes Java Applet, das eine komfortable Web-Oberfläche für den ContentBroker zur Verfügung stellt. Zur Benutzung ist ein registrierter User erforderlich, je einer für jede einliefernde Institution (Contractor).

Mit Hilfe dieser Anwendung lässt sich etwa der Verarbeitungsstatus eingelieferter Pakete überwachen oder ein bereits vollständig archiviertes Pakets anfordern (Retrieval). Hierzu fordert der ContentBroker ein bestimmtes Datenpaket von iRODS an. Der logische Pfad, bzw die URN des Objekts genügt hierzu, alle weiteren Aufgaben übernimmt iRODS intern.

Das angeforderte Paket wird in einem dem Benutzer zugeordneten Ordner bereitgestellt und kann über die DA Web Oberfläche heruntergeladen werden. In Zukunft ist geplant, DA-Web mit einer Web-DAV Schnittstelle auszurüsten, um den Ingest hierüber abzuwickeln zu können.

FAZIT: DA-Web gewährt Zugriff auf die vollständige Sammlung eines (bei DA-NRW institutionellen) Kontraktors. Im Hinblick auf die Anforderung eines auf Individuen ausgelegten Access Management stellt dies insofern kein Problem dar, wenn jedes Individuum als einliefernder Kontraktor betrachtet werden kann. Andernfalls wären hier weitere Mechanismen zur Zugriffskontrolle zu implementieren.

DAS TESTSYSTEM

Das auf einer Virtuellen Maschine (nighthorse09.dai-cloud.uni-koeln.de) aufgesetzte Testsystem verfügt über alle zuvor genannten Komponenten und besteht derzeit aus zwei „Knoten" zur Datenhaltung.

Fedora wurde in Version 3.7 installiert, jedoch wurde während der weiteren Entwicklung von DA-NRW ein Downgrade auf 3.5 vorgenommen, da Fedora mit der letzten Version neue Mechanismen zur Steuerung des Zugriffs vorgenommen hat.

LINKS

DA-Web: http://nighthorse09.dai-cloud.uni-koeln.de:8080/daweb3/

Fedora: http://nighthorse09.dai-cloud.uni-koeln.de:8080/fedora/

(Passwörter bei Bedarf über ianus-fdz@dainst.de)