New Comer Forum

Maria Lusky*

Entwicklung einer Callimachus-Anwendung zur Datenpflege im Digitalen Archiv der Pina Bausch Foundation

DOI 10.1515/iwp-2016-0006

Zusammenfassung: In diesem Beitrag wird die Entwicklung einer Anwendung zur Pflege einzelner Datensätze im Digitalen Archiv der Pina Bausch Foundation nachgezeichnet. Zunächst werden die Anforderungen des Linked Data-Archivs angeführt. Als verwendetes Werkzeug zur Anwendungsentwicklung wird das Linked Data Management System Callimachus vorgestellt, woraufhin einzelne Aspekte der Entwicklung erläutert werden. Nach einer kritischen Bewertung folgt eine Reflexion über Callimachus als Werkzeug für Informationswissenschaftler.

Deskriptoren: Prototyp, Dokumentation, Multimedial, Datenverknüpfung, Datenstruktur, Informationsnetz, Archiv, Digital, Software, Callimachus, Berufsbild, Informationswissenschaft

Developing an application with Callimachus for data maintenance in the Digital Archive of the Pina Bausch Foundation

Abstract: This paper presents the development of an application for data maintenance in the Digital Archive of the Pina Bausch Foundation. First, the requirements of the Linked Data archive are outlined. The Linked Data Management System Callimachus is introduced, whereupon aspects of the approach of the application development are described. In the last part Callimachus is evaluated and discussed as a tool for information scientists.

Descriptors: Prototype, Documentation, Multimedia, Data capturing, Data structure, Data Linkage, Archive, Digital, Software, Callimachus

Développement d'une application avec Callimachus pour la gestion des données de l'archive numérique de la Fondation Pina Bausch

Résumé: Ce document présente le développement d'une application pour la gestion des données de l'archive numérique de la Fondation Pina Bausch. En premier lieu, les critères spécifiques à l'archive Linked Data sont définis. Le système de gestion Linked Data Callimachus est présenté, puis les différents aspects relatifs à l'approche du développement de l'application sont décrits. Dans la dernière partie, Callimachus est évalué et le thème de son utilisation comme outils d'information à l'intention des chercheurs est abordé.

Descripteurs: Documentation, Multimédia, Gestion des données numériques, Structure des données numériques, Hyperlink, Archive numérique, Logiciel, Callimachus

Das Digitale Archiv

Die Pina Bausch Foundation hat es sich zur Aufgabe gemacht, das kulturelle Vermächtnis der Tänzerin und Choreografin Pina Bausch zu bewahren, lebendig zu halten und Menschen nahezubringen.¹ Neben Digitalisaten und Metadaten unter anderem zu Musik, Fotografien, Kostümen, Stücken, Aufführungen und Personen wird auch der Prozess der Stückentstehung im digitalen Archiv dokumentiert.

Das Digitale Archiv wurde von 2011 bis Ende 2013 im Rahmen einer Forschungskooperation zwischen der Pina Bausch Foundation und dem Institut für Kommunikation und Medien (IKuM) an der Hochschule Darmstadt begonnen. Bei der Bestandsentwicklung wurde zunächst die Unterstützung des Digitalisierungsprozesses in den Vordergrund gestellt, was zu großen Datenmengen führte, die

*Kontaktperson: Maria Lusky, Hochschule Darmstadt, Informationswissenschaft, Pina Bausch Foundation, Fingscheid 5, 42285 Wuppertal, E-Mail: maria.lusky@gmail.com

¹ Pina Bausch Foundation (Hg.) (o. J.)

als Linked Data in einen Sesame Triple Store übertragen wurden.

Die Generierung dieser Daten als Linked Data erfolgte im digitalen Archiv zunächst über Excel-Spreadsheets, deren Inhalt mittels mehrerer Skripte in Linked Data überführt und schließlich im Triple Store abgespeichert wurde. Die Daten eines Excel-Spreadsheets bilden dadurch einen Kontext, über den die Herkunft der Daten nachvollzogen werden kann. Besonders problematisch sind hierbei einerseits die hohe Fehleranfälligkeit der Daten beim Eintragen in die Spreadsheets, andererseits der Aufwand einer Korrektur: Um einen Fehler zu beheben, muss das betroffene Excel-Spreadsheet ausfindig gemacht, manuell korrigiert und erneut durch das Skript verarbeitet werden.

Mit dem Wachstum des digitalen Archivs verlagerte sich der Bedarf des schnellen Inputs großer Datenmengen hin zum gezielten Erstellen und Korrigieren weniger Datensätze, wodurch der Einsatz eines nativen Frontend zur interaktiven Pflege der Ressourcen als Ablöse der Excel-Spreadsheets unabdingbar wurde. Im Rahmen einer Bachelorarbeit an der Hochschule Darmstadt wurde in Zusammenarbeit mit der Pina Bausch Foundation eine prototypische Anwendung mit Callimachus entwickelt.

Konzeption und Entwicklung

Anforderungen

Aufbau und Pflege eines Linked Data Archivs allgemein und des Digitalen Archivs der Pina Bausch Foundation im Besonderen stellen bestimmte Ansprüche an ein Data Management System. Zusätzlich zur Berücksichtigung der Nutzerbedürfnisse sollte es sich an den Eigenschaften eines Linked Data-Archivs orientieren. Neben vielen sehr speziellen Anforderungen an das Frontend gibt es auch einige grundlegende notwendige Features.

Wichtigste Funktion des Systems ist die Fehlerreduktion, welche einerseits in der einfachen Korrektur bereits vorhandener Fehler, andererseits in der Vermeidung fehlerhafter Dateneingaben besteht und eine Erhöhung der Datenqualität zur Folge hat.

Bei der bisherigen Arbeit am digitalen Archiv führte besonders die manuelle Vergabe der Uniform Resource Identifiers (URIs) zu Problemen. Der Aufbau der URIs folgt einem bestimmten im Rahmen des digitalen Pina Bausch-Archivs gültigen Schema. Die Zuordnung zu den einzelnen Ressourcen geschieht durch manuelles Eintippen eines URIs in ein Excel-Spreadsheet. Diese Methode ist äußerst fehleranfällig und führt zu nicht verknüpften Datensätzen und falsch abgelegten, nicht aufrufbaren Daten. Eine Auf-

gabe der neuen Software ist daher die automatisierte Generierung von URIs nach dem Schema des digitalen Archivs.

Hilfreich ist zudem eine Versionsgeschichte, die Aufschluss darüber gibt, welche Änderungen wann und von wem an einer Ressource vorgenommen wurden. Bezieht sich eine solche Versionsgeschichte auf einzelne Ressourcen statt auf ganze Kontexte (Excel-Spreadsheets), wird Transparenz hinsichtlich der Datenbearbeitung geschaffen und eine gezielte Revision ermöglicht.

Eine weitere Anforderung ist die Offenheit, es jedem Nutzer zu ermöglichen, eine Ressource zu annotieren oder zu kommentieren.

Zusätzlich dazu sollte das Data Management System über eine individualisierbare Rechteverteilung verfügen, die verschiedene Nutzergruppen unterscheidet.

Da die Stiftung in der Vergangenheit keine und zum gegenwärtigen Zeitpunkt nur eine sehr kleine IT-Abteilung unterhält, sollte der Pflegeaufwand des Systems möglichst gering ausfallen. Insbesondere müssen eine bedarfsweise Anpassung an das Datenmodell und ein entsprechender Grad an Flexibilität des Systems gewährleistet sein, weil das Datenmodell mit fortschreitendem Wachstum des Archivs ebenfalls erweitert wird. Auch die Eingliederung in die bereits bestehende Systemarchitektur und die Anbindung an den Triple Store zum Datenimport und -export sollten mit einmaligem Aufwand umsetzbar sein.

Das Linked Data Management System Callimachus

Um eine Anwendung zu entwickeln, welche den genannten Anforderungen gerecht wird, wurde das Linked Data Management System Callimachus² verwendet. In Callimachus können Anwendungen zum Erstellen, Bearbeiten, Speichern und Verwalten von Linked Data erstellt werden. Es umfasst einen internen Triple Store, in dem Daten wahlweise direkt gespeichert oder über eine REST API oder einen SPARQL Endpoint importiert werden können. Callimachus selbst und die damit erstellten Anwendungen sind nach einem Wiki-Prinzip aufgebaut: Jede Entität wird mit ihren Eigenschaften auf einer Seite angezeigt. Über eine Navigation hat ein Nutzer die Möglichkeit, sie zu bearbeiten oder auf der Diskussionsseite dieser Entität einen Kommentar dazu zu verfassen, während jede Änderung und jeder Beitrag in der ebenfalls einsehbaren Versionsgeschichte

² Weitere Informationen unter <http://callimachusproject.org>.

dokumentiert werden. Darüber hinaus bietet Callimachus die Möglichkeit, Nutzer über ein Rechtesystem zu verwalten. Die Informationen zu jeder Entität können im Turtle-³, RDF/XML- oder JavaScript Object Notation for Linked Data (JSON-LD)-Format heruntergeladen werden.

Das in Callimachus enthaltene Rechtesystem umfasst fünf Gruppen, die bestimmte Rechte repräsentieren und denen die Nutzer zuzuordnen sind⁴. Es kann bei Bedarf noch erweitert werden. Die Zuweisung der Rechte erfolgt global, für einzelne Klassen oder auch nur für bestimmte Funktionen einzelner Klassen.

Kern einer Callimachus-Anwendung bilden die Klassen, deren Entitäten mit der Anwendung bearbeitet werden. Jede Klasse wird mit Seiten-Templates zum Visualisieren, Bearbeiten und Erstellen ihrer Ressourcen verknüpft. Bei den Templates zum Erstellen und Ändern von Ressourcen handelt es sich um XHTML-Formulare mit eingebettetem RDF in attributes (RDFa), welches in Kombination mit spezieller Callimachus-Syntax Statements der Abfragesprache Sparql Protocol And RDF Query Language (SPARQL) erzeugt, mittels derer die Daten als Tripel erfasst und im Triple Store abgespeichert werden. Templates zur Visualisierung sind in ihrer einfachsten Form XHTML-Seiten, in die Daten über RDFa eingelesen werden. Über mit XHTML-Seiten verknüpfte SPARQL-Queries werden Suchergebnisse als Listen angezeigt und in andere Templates eingebunden. Bei Bedarf können über den Standard von Callimachus hinaus gehende Funktionen durch JavaScript-Files umgesetzt werden. Als Start- oder Zwischenseiten der Anwendung dienen Seiten im Docbook-Format.

Weiterhin können Callimachus-Anwendungen durch XProc-Pipelines, XQuery-, XInclude und XSL Transformation (XSLT)-Files ausgebaut werden. Diese wurden im Rahmen dieser Anwendungsentwicklung allerdings nicht verwendet.

Entwicklung des Prototypen

Da eine Callimachus-Anwendung auf Klassen und den mit ihnen verknüpften Templates aufbaut, wurden zunächst die im Digitalen Archiv der Pina Bausch Foundation verwendeten Ontologien in Callimachus geladen und die darin enthaltenen relevanten Klassen ausgewählt. Anschließend wurde jede dieser Klassen mit jeweils drei

Templates zum Visualisieren, Bearbeiten und Erstellen von Datensätzen versehen. Die in Callimachus vorgegebenen Templates mussten um die Eigenschaften der jeweiligen Klasse ergänzt werden, indem für jede Property ein XHTML- und RDFa-Codeabschnitt eingefügt wurde.

Callimachus bildet URIs standardmäßig aus der über `rdfs:label` festgelegten Benennung der Ressource, während die URIs des Pina Bausch Archivs einem komplexeren Schema entsprechen. Da Callimachus JavaScript für die Einbindung eigener Funktionen vorsieht, wurden JavaScript-Funktionen geschrieben, um die URIs neuer Ressourcen nach dem Schema des digitalen Pina Bausch Archivs zu erzeugen. Einfache HTML5-Attribute, beispielsweise für den Abgleich von Eingaben mit regulären Ausdrücken, und Möglichkeiten der Callimachus-Syntax wie das Einlesen der Ergebnisse von SPARQL-Queries über XInclude in Dropdown-Menüs beugen Fehlern beim Pflegen der Daten in der Anwendung vor und tragen maßgeblich zur Erhöhung der Datenqualität bei.⁵ Die Kommentarfunktion ermöglicht das Annotieren von Entitäten und fördert die Kommunikation über Änderungen zwischen den Nutzern. Die Informationen der Versionsgeschichte ermöglichen einen im Vergleich zu den Excel-Spreadsheets deutlich transparenteren und effizienteren Gebrauch der Historie auf Ebene einzelner Ressourcen statt ganzer Daten-Kontexte, sodass beim Bearbeiten der Daten in Callimachus die Schritte des Suchens und erneuten Uploads gespart werden.

Weniger einfach gestaltet sich der Datenaustausch zwischen dem Callimachus Store und dem derzeit vom digitalen Archiv genutzten Sesame Triple Store. Der Callimachus Triple Store zieht keine Inferenzen, welche für das Digitale Archiv der Pina Bausch Foundation hingegen essenziell sind und rund die Hälfte des Datenvolumens ausmachen.⁶ Aus diesem Grund wird zur Speicherung der Daten statt des Callimachus Stores weiterhin der Sesame Triple Store genutzt, was wiederum eine Export- und Import-Lösung erfordert. Dazu wird die Callimachus REST API verwendet, über die Daten regelmäßig zwischen den Stores ausgetauscht werden können.

Die Anwendung ermöglicht zurzeit die Bearbeitung von Ressourcen aus neun Klassen, darunter Stücke, Aufführungen, Personen, Bühnenbilder sowie digitalisierte und physische Videos. Für jede dieser Klassen wurde mittels der drei Templates ein kompletter Interaktionszyklus für die Operationen Create, Read, Update und Delete (CRUD) erstellt. Im Zuge der Nutzung kann man

³ Weitere Informationen zur Terse RDF Triple Language unter <http://www.w3.org/TR/turtle/>.

⁴ 3 Round Stones (Hg.) (2015).

⁵ Luský (2014) S. 36–37.

⁶ Thull (2014).

Abbildung 1: Bearbeitung einer Aufführung in der Callimachus-Anwendung: Es können weitere Informationen, beispielsweise zu Besetzung und Orten, hinzugefügt werden (Darstellung: Maria Lusky).

über die Klassenhierarchie oder über Verknüpfungen der Ressourcen untereinander zu einer Ressource navigieren. Um eine bestimmte Ressource aufzurufen, kann man auch die Suchfunktion nutzen und dadurch direkt zu ihr gelangen.

Bewertung des Lösungsansatzes

Die Entwicklung einer prototypischen Anwendung mit Callimachus wurde als Machbarkeitsstudie im Rahmen einer Bachelorarbeit durchgeführt. Daher sollten die Möglichkeiten und Grenzen Callimachus' kritisch beleuchtet werden.

Zunächst ist festzustellen, dass Callimachus durch die Formulare und darauf basierender Plausibilitätsprüfung sehr gut zur Reduktion und Korrektur von Eingabefehlern geeignet ist und somit maßgeblich zur Erhöhung der Datenqualität im digitalen Archiv beiträgt (Abb. 1). Die automatisierte Generierung der URIs ist ebenfalls gegeben und ließ sich durch JavaScript-Funktionen an das URI-Schema des digitalen Archivs anpassen. Die Historie einzelner Ressourcen, ein individualisierbares Rechtssystem sowie die Kommentarfunktion sind durch die

Wiki-Struktur bereits in Callimachus enthalten und kommen der Arbeit am digitalen Archiv nachhaltig zu Gute.⁷

Ein Nachteil ist hingegen das Scripting der Properties in den Seiten-Templates. Jede Property muss im XHTML-Code angelegt sein, was zu einer unflexiblen Anwendung führt und bei Änderung des Datenmodells eine manuelle Aufarbeitung der Templates erforderlich macht. Da Callimachus ständig weiterentwickelt wird, kann es passieren, dass ein Update nicht nur neue Funktionen bereit stellt, sondern auch einige Code-Patterns nicht mehr unterstützt, was wiederum dazu führt, dass die betreffenden Stellen im Code der Anwendung neu überarbeitet werden müssen. Dem Anspruch des geringen Pflegeaufwands wird Callimachus daher weniger gerecht.

Die Integration von Callimachus in die bereits bestehende Systemarchitektur erwies sich als aufwendiger als erwartet, was insbesondere darauf zurück zu führen ist, dass der Callimachus Triple Store keine Inferenzen ziehen kann. Für den Datenbestand des digitalen Pina Bausch-Archivs, in dem Inferenzen eine große Rolle spielen, ist

⁷ Lusky (2014) S. 46.

dies ein Problem, welches über zusätzlichen Aufwand unter Einsatz der REST API gelöst werden musste.⁸

Eine weitere Schwäche von Callimachus ist die oberflächlich abgefasste und wenig umfangreiche Dokumentation, die umfassendes Grundwissen aus Bereichen der Informatik voraussetzt, dadurch zu einer vergleichsweise langen Einarbeitungszeit führt und das Entwickeln von komplexeren Callimachus-Anwendungen sehr aufwendig macht.

Die bereits erwähnte stetige Weiterentwicklung ist gleichzeitig auch eine große Stärke Callimachus': Mit neuen Funktionen können Anwendungen genauer auf den Bedarf abgestimmt werden. Auch im Prozess der Entwicklung der hier thematisierten Anwendung kamen durch Updates noch hilfreiche Features hinzu. Callimachus zeichnet sich weiterhin durch sehr guten Service seitens seiner Entwickler aus, vom dem aufgrund der nicht immer hilfreichen Dokumentation gerne Gebrauch gemacht wurde. In einer Google Group, per E-Mail oder Chat werden Fragen stets zeitnah beantwortet.

Den Vor- und Nachteilen des Systems stehen wiederum die für die Anwendungsentwicklung benötigten Kenntnisse gegenüber. Zur Erstellung der Templates und der Funktionen der Anwendung müssen hauptsächlich HTML5, Cascading Style Sheets (CSS) und JavaScript mit jQuery beherrscht werden, weiterhin verschiedene XML-basierte Standards wie XSLT, XQuery, XProc und XInclude. Für die inhaltliche Ebene der Anwendung und die Daten selbst werden umfassende Kenntnisse aus dem Bereich Linked Data, insbesondere Turtle, RDFa und SPARQL benötigt. RDF, RDF-Schema, die Web Ontology Language (OWL) und das Simple Knowledge Organization System (SKOS) werden vorausgesetzt. Bei der hier behandelten Anwendung wurden zudem die Pina Bausch Archive Ontology, Functional Requirements for Bibliographic Records (FRBR), Dublin Core sowie Friend Of A Friend (FOAF) und das PROV-Datenmodell verwendet. Um Callimachus auf einem Server zu installieren und um den Datenimport und -export umzusetzen, werden zudem Grundkenntnisse des jeweiligen Servers und der REST-API benötigt.

Zusammenfassend ist festzustellen, dass Callimachus ein Werkzeug mit Potenzial ist, um Linked Data zu verwalten, welches für komplexere Anwendungen allerdings eine zeitaufwendige Einarbeitung verlangt.

Diskussion

Im Zusammenhang mit der Entwicklung des Prototypen sind verschiedene Tätigkeiten angefallen, von denen ein Teil nicht zur Kernkompetenz eines Informationswissenschaftlers gehört. Daher soll die Eignung von Callimachus als Werkzeug für Informationswissenschaftler diskutiert werden.

Die Konfiguration von Callimachus, die Konzeption der Anwendung und das Skripting unter Verwendung von HTML, JavaScript und XML-basierten Standards sowie Kenntnisse im Bereich Linked Data sind Inhalte, die der Schnittmenge von Informationswissenschaft und Informatik zugeordnet werden können.

Typische informationswissenschaftliche Kompetenzen sind die Einarbeitung und Pflege der Archivdaten als Linked Data sowie die Datenmodellierung. Beide Tätigkeiten finden jedoch vor bzw. nach der Anwendungsentwicklung statt und tragen daher nicht direkt dazu bei. Arbeitsschritte der klassischen Informationswissenschaft sind daher eher im Bereich der Nutzer der Callimachus-Anwendung einzuordnen. Weiterhin wurden Usability-Tests zur Gestaltung des User Interfaces durchgeführt. Eine anwenderzentrierte Entwicklung wurde angestrebt, fand ihre Grenzen allerdings in einigen Fällen durch die von Callimachus vorgegebene Informationsarchitektur.

Für die hier beschriebene Entwicklung eines Prototypen wurden Kenntnisse der Systementwicklung, die über das typische Qualifikationsprofil der Informationswissenschaft hinausgehen, selbständig erlernt. Die Weiterbildung erfolgte gezielt und war dennoch zeitintensiv. Obwohl die Anwendungsentwicklung im hier vorgestellten Projekt erfolgreich war, komme ich daher zu dem Schluss, dass eine derartige Entwicklung ohne Kenntnisse der Systementwicklung außerhalb einer Forschungsarbeit als unwirtschaftlich bezeichnet werden kann.

Perspektive

Mit dem Linked Data Management System Callimachus wurde ein Prototyp zu Datenaufbau und -pflege im Digitalen Archiv der Pina Bausch Foundation entwickelt. Es wurde auch gezeigt, dass eine Anwendungsentwicklung in Callimachus mit Mitteln der Informationswissenschaft durchgeführt werden kann, wenn Informatik-Inhalte zusätzlich erlernt wurden oder werden.

Ein großer Vorteil bei der Entwicklung einer Anwendung besteht in der Zusammenarbeit von Informationswissenschaftlern und Informatikern dann, wenn die Backend-Programmierung von Informatikern abgedeckt wird und

⁸ Lusky (2014) S. 45.

Informationswissenschaftler Kenntnisse aus der Nutzerforschung, beispielsweise aus Information Behavior⁹, User Experience und Usability, einbringen.

Danksagung: Besonderer Dank gilt meiner Kollegin Vera Marz von der Pina Bausch Foundation für die konstruktive Zusammenarbeit, die Unterstützung bei Fragen zum Archiv und die Bereitstellung der Datengrundlage. Dank geht ebenso an Prof. Dr. Bernhard Thull sowie an Kerstin Diwisch für ihre Hilfe in Fragen der Systementwicklung.

Literatur

3 Round Stones (Hg.) (2015): Callimachus for Web Developers. Online verfügbar unter <http://callimachusproject.org/docs/1.3/callimachus-for-web-developers.docbook?view>, zuletzt aktualisiert am 14.01.2015 [25.11.2015].

Lusky, Maria (2014): Anwendungsentwicklung mit dem Linked Data Framework Callimachus am Beispiel des digitalen Archivs der Pina Bausch Stiftung. Bachelorarbeit. Hochschule Darmstadt, Darmstadt.

Pina Bausch Foundation (Hg.) (o. J.): Mission. Online verfügbar unter <http://www.pinabausch.org/de/foundation/mission> [25.11.2015].

Thull, Bernhard (2014): Das Digitale Pina Bausch Archiv. In: Marc Wagenbach und Pina Bausch Foundation (Hg.): Tanz Erben. Pina lädt ein. Bielefeld: Transcript, S. 59–73.

Maria Lusky
Hochschule Darmstadt
Informationswissenschaft

Pina Bausch Foundation
Fingscheid 5
42285 Wuppertal
maria.lusky@gmail.com

⁹ Die Bedeutung dieses Themenbereichs für die Informationswissenschaft wurde in „Information. Wissenschaft und Praxis“ Band 66, Heft 1 vom Februar 2015 thematisiert.