Dokumentation, Werkzeugkasten, Pakete - Nachhaltigkeit von Daten und Funktionalität Digitaler Editionen

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Die Rolle digitaler Ressourcen in den Geisteswissenschaften und die zunehmende Bedeutung von Algorithmen werden noch immer unterschätzt. In den letzten Jahren bzw. Jahrzenten wurden zahlreiche Software-Werkzeuge, virtuelle Forschungsumgebungen und interaktive Publikationen, wie Datenbanken oder Digitale Editionen, für die geisteswissenschaftliche Forschung entwickelt. Diese werden innerhalb der Forschungscommunity mit zunehmender Tendenz akzeptiert und inzwischen breit eingesetzt. Jedoch stehen wir heute vor der Herausforderung diese verschiedenen Ressourcen, die zu einem großen Teil auf unterschiedlichen technischen Grundlagen basieren, weiter zu pflegen und verfügbar zu halten. Transparenz und Reproduzierbarkeit von Forschungsergebnissen, die mit einer digitalen Ressource oder Software erstellt wurden leiden darunter, dass diese Software oft nach wenigen Jahren nicht mehr gepflegt wird und damit nicht mehr lauffähig ist.

Der Vortrag beleuchtet die hier skizzierte Problematik am Beispiel Digitaler Editionen. Es werden drei Punkte vorgeschlagen, wie durch einen Bottom-Up-Ansatz die Nachhaltigkeit digitaler Ressourcen gefördert werden kann. Der Fokus liegt dabei auf den Erfahrungen, die im Laufe der letzten 10 Jahre mit der Entwicklung von XML-basierten Digitalen Editionen und dem Einsatz ausgewählter Software-Werkzeuge, wie der nativen XML-Datenbank eXistdb (http://exist-db.org), gewonnen wurden.

Bei digitalen Ressourcen handelt es sich um dynamische Objekte. Das bedeutet einerseits, dass die Inhalte jederzeit korrigiert, erweitert oder verändert werden können. Andererseits muss die technologische Basis fortlaufend aktuell, sicher und verfügbar gehalten werden. Diese beiden Prozesse sind Teilaufgaben eines größeren Aufgabenbereichs, der unter dem Begriff „ *data curation*“ zusammengefasst werden kann.

Eine Digitale Edition ist mehr als nur ihre Forschungsdaten. Letztere, in vielen Fällen XML-Dokumente, werden oft erst durch eine adäquate Darstellung, durch Visualisierungen, wie Text-Bild-Verlinkungen, verschiedene Ansichten auf den Text, Netzwerke oder Timelines oder auch Verknüpfungen mit anderen externen Ressourcen „zum Leben erweckt“. Dieses Leben in Form programmierter Funktionalität ist ein genuiner Forschungsbereich der Digital Humanities. Allerdings führt die Funktionalitätsschicht (siehe Abbildung 1) einer Digitalen Edition auch dazu, dass *data curation* eine sehr komplexe Aufgabe werden kann, da hier die Flexibilität der Implementierung erheblich höher ist als auf der Ebene der Datenschicht.

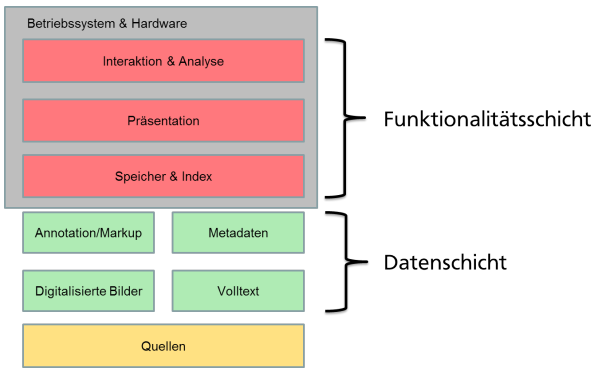

Abbildung 1: Schichtenmodell Digitale Edition

Digitale Editionen sind Software-Werkzeuge für die Analyse von Forschungsdaten. Damit sind sie ein Teil des Forschungsprozesses, der erhalten werden muss, um die Reproduzierbarkeit der Forschungsergebnisse zu gewährleisten.

Eine digitale geisteswissenschaftliche Ressource durchläuft üblicher Weise einen typischen Lebenszyklus. Dieser beginnt mit der Analyse der analogen Quellen, geht über die Datenmodellierung, die Auswahl bzw. Anpassung oder Neuentwicklung von Bearbeitungswerkzeugen sowie der digitalen Publikation der Forschungsergebnisse bis hin zu Fragen der Langzeitverfügbarkeit und Langzeitarchivierung. Bei jedem dieser Schritte sind verschiedenen Kompetenzen involviert. Das bedeutet, der Aufbau einer Digitalen Edition ist immer Teamwork. Dieses Team setzt sich in den meisten Fällen aus Personen zusammen, die einerseits das inhaltliche Fachwissen mitbringen und andererseits aus Personen mit einer Vielzahl unterschiedlicher technischer Kompetenzen:

1. Analyse der Quellen (Geisteswissenschaftler)
2. Anforderungsanalyse der digitalen Ressource, *Requirement Engineering* (alle Projekbeteiligten)
3. Entwurf des Daten- / Dokumentenmodells, Auswahl von Standards (Geisteswissenschaftler, Datenbankspezialisten, Markupspezialisten, Metadatenspezialisten)
4. Auswahl, Anpassung bzw. Entwicklung von Tools (Programmierer, Geisteswissenschaftler)
5. Aufsetzen und Betreuen der Server (Systemadministratoren)
6. Konzept, Design und Umsetzung der Web-Publikation (Webdesigner, Webentwickler, Geisteswissenschaftler)

7. Vorbereitung für Langzeitverfügbarkeit / -archivierung (Metadatenspezialisten, Dokumentationsspezialisten)
8. Betreuung und Wartung nach Projektende („ *data curators*“)

An jeder Stelle in diesem Lebenszyklus einer digitalen Ressource werden Entscheidungen getroffen, die Auswirkungen auf den nachfolgenden Schritt haben. So bilden die eigentliche Analyse der Inhalte, die z.B. in einer Digitalen Edition publiziert werden sollen, und die Anforderungsanalyse das Fundament, auf dem alles aufbaut, vom Daten- oder Dokumentenmodell, bis hin zur Publikation und der *data curation*.

Aus methodischer Sicht, mit besonderem Augenmerk auf das zugrundeliegende Text- bzw. Dokumentenmodell, wurden Digitale Editionen bereits ausführlich beschrieben (siehe Pierazzo 2015 und Sahle 2013). Eine Analyse aus technischer Sicht steht noch aus. Um die Entwicklung, Betreuung und Nachhaltigkeit Digitaler Editionen zu gewährleisten bedarf es eines technologischen Publikationskonzepts, das aus möglichst standardisierten Komponenten besteht.

Bisher existieren sehr erfolgreiche Standardisierungen auf dem Gebiet der Metadaten und der Textauszeichnungen, z.B. mit den Richtlinien der *Text Encoding Initiative* (TEI), aber wenig bis gar nichts bei der technischen Umsetzung und der Dokumentation. Dies würde helfen anschlussfähigere, stabilere und nachhaltigere digitale Ressourcen aufzubauen und damit auch die Arbeit eines Datenkurators, der sich um die Pflege dieser Ressourcen nach Projektende kümmert, deutlich vereinfachen.

Für eine aussichtsreiche Nachhaltigkeit kann die Lösung nicht allumfassend sein, sondern nur für klar definierte Anwendungsfälle gelten. In dem hier vorgestellten Fall ist dies eine XML-basierte Digitale Edition, die mit Technologien aus der X-Familie (XSLT, XQuery, XML-Schema, eXistdb) entwickelt wird. Das Grundprinzip ist jedoch auf andere Anwendungsszenarien übertragbar:

1. Eine ausführliche Dokumentation
2. Ein klar definierter Werkzeugkasten
3. Die Paketierung aller Projektressourcen

Dokumentation

Eine nachhaltige digitale Forschungsressource bzw. –software ist langfristig verfügbar, gut dokumentiert, lizenziert und versioniert, um die Reproduzierbarkeit der Forschungsprozesse zu garantieren. Die wichtigste Komponente ist eine ausführliche, formalisierte Dokumentation, die mindestens die folgenden Informationen enthalten sollte:

- Den Namen des Projekts und aller beteiligten Institutionen und Personen.
- Den Projektstatus: geplant, in Arbeit, veröffentlicht, beendet.
- Die eingesetzten Technologien und Standards inklusive Versionsangabe.
- Lizenzangaben zu Forschungsdaten, Quellcode, und anderen Komponenten, wie Schriftarten, Audio- oder Videodokumenten.
- Informationen darüber, wo der Quellcode und die Forschungsdaten zu finden sind.
- Informationen über die bereitgestellten APIs und andere Schnittstellen, um die Forschungsdaten in verschiedenen Formaten abzurufen (XML, HTML, PDF, JSON usw.) und in anderen Kontexten weiterzuverarbeiten.
- Details über die Forschungsmethode und den Hintergrund des Projekts. (Mehr dazu siehe Faniel 2015)
- Zitations- und Referenzierungsanweisungen für die persistente Adressierung aktueller und älterer Versionen der Forschungsdaten, Metadaten und Software.
- Eine standardisierte Historie der Projektentwicklung.

Selbstverständlich kann diese Liste nur ein erster Vorschlag sein. Sie enthält keinesfalls alle möglichen Informationen, die zu einer Digitalen Edition angegeben werden können. Die Dokumentation sollte maschinenlesbar (um z.B. als XML oder JSON weiter verarbeitet werden zu können) und über eine standardisierte Adresse bzw. einen klar definierten Zugriffspunkt (z.B. http://home.of.project/api/projectdescription) abrufbar sein. Dadurch wäre es möglich eine Digitale Edition bei einem zentralen Verzeichnis anzumelden, in dem alle Informationen und Updates über Digitale Editionen, die demselben Publikationsmodell folgen, gesammelt werden. Ein solches Verzeichnis existiert noch nicht.

Definierter Werkzeugkasten

Wie oben beschrieben kann Nachhaltigkeit nur in einem definierten Rahmen hergestellt werden, indem man klare Anwendungsfälle beschreibt. Selbst in diesen sind die Möglichkeiten der Umsetzung nahezu unbegrenzt. Daher ist es wichtig, genau zu definieren, welche Technologien, Standards und Software-Werkzeuge zum Einsatzkommen und welche Abhängigkeiten bestehen. Es ist ratsam die Zahl der eingesetzten Tools überschaubar zu halten und sich auf etablierte und gut dokumentierte Technologien zu konzentrieren.

Paketierung

Alle zusammengehörenden Komponenten einer Digitalen Ressource (Daten, Metadaten, Quellcode, Binärdatein, Dokumentation) müssen immer zusammen abrufbar sein. Diese Pakete tragen deutlich zu einer nachhaltigeren Entwicklung bei. Es ist immer klar, wo sich alle relevanten Informationen und Daten befinden. Zudem könnte ein Paket einer zentralen Kurationsstelle, z.B. einem Digital Huamnities Data Center, übergeben werden, die sich um die Betreuung der digitalen Ressourcen abgeschlossener Projekte kümmert. Diese Anlaufstelle existiert ebenfalls noch nicht.

Für den hier vorgestellten Anwendungsfall bietet das von eXistdb verwendete EXPath-Format (http://

expath.org/) einen guten Ausgangspunkt für die Paketierung. Dieses Format beschreibt ein Packaging-System (http://expath.org/modules/pkg/), das es erlaubt XML-basierte Dokumente zusammen mit Abfrage- und Transformationsskripten sowie verschiedenen anderen Ressourcen auf eine standardisierte Art und Weise zu paketieren, das dieses Paket von allen Softwaresystemen, die diesem Standard folgen verstanden und ausgewertet werden kann. Damit kann dieses Packagesystem ähnlich fungieren, wie ein App-Store für Smartphones. Auch der Anpassungsaufwand für Software, die während der Projektlaufzeit eingesetzt wird, würde sich so verringern.

Andere Anwendungsszenarien lassen sich nicht auf Softwareebene paketieren. In diesen Fällen kann man auf Anwendungsvirtualisierung zurückgreifen, wie sie zum Beispiel mittels Docker (https://www.docker.com/) möglich ist.

Um eine solche Standardisierung auf der technischen Ebene durchzuführen benötigt es eine aktive Digital-Humanities-Enwicklercommunity. Die Rolle des DH-Entwicklers ist im allgemeinen Diskurs noch deutlich unterrepräsentiert. Um wirklich erfolgreich Softwaretools für die geisteswissenschaftliche Forschung programmieren zu können, benötigt ein Entwickler mehr als nur ein grundlegendes Verständnis der typischen Problematiken in den einzelnen Fachdisziplinen. Umgekehrt erlauben Programmierkenntnisse ein wesentlich besseres Verständnis über die Funktionsweise der Software und damit bessere Einsatzmöglichkeiten sowie das Potential zu eigenen Verbesserungen. Um hier Fortschritte zu erzielen, müssen sich die DH-Entwickler besser organisieren. Die DH-Entwicklercommunity wäre der Kreis an Personen, die die Einführung und Anwendung von Standards, wie dem EXPath-System diskutieren und tragen. Mittelfristig wäre das Ziel für jeden Schritt im Lebenszyklus einer digitalen Ressource einen oder mehrere Vorschläge einer standardisierten Herangehensweise in Form eines *best-practice*-Leitfadens zu haben, der sich als *technical reader* zur Erstellung digitaler geisteswissenschaftlicher Ressourcen eignet und Vorschläge zu Schnittstellen, Standards, Lizenzen, Dokumentation, Zitationshinweise usw. anbietet.

Ein Ansatzpunkt dafür bietet die Arbeit des 2010 gegründeten Software Sustainability Institute (https://www.software.ac.uk/), Hong 2010 sowie Hettrick 2016, die verschiedene Ansätze zur Forschungs-Software-Nachhaltigkeit untersucht haben.

Bibliographie

Faniel, Ixchel (2015): *Data Management and Curation in 21st Century Archives*, 21. September 2015, http://hangingtogether.org/?p=5375 .

Hettrick, Simon (2016): *Research Software Sustainability. Report on a Knowledge Exchange Workshop*.

Hong, N. Chue et al. (2010): *Software Preservation Benefits Framework. Software Sustainability Institute Technical Report*.

Pierazzo, Elena (2015): *Digital Scholarly Editing, Theories, Models and Methods*. Ashgate.

Sahle, Patrick (2013): *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. 3 Bände. Norderstedt: Books on Demand.