Disseminations-Services: dhPLUS

Hörmann, Richard; richard.hoermann@sbg.ac.at

dhPLUS ist eine Plattform der Universität Salzburg (Paris Lodron Universität Salzburg: PLUS), die die Langzeitarchivierung (LZA) von Digital Humanities(DH)-Projekten sicherstellen soll. Sie ist vorzugsweise für Projekte der Universität Salzburg gedacht, steht aber darüber hinaus der nationalen und internationalen DH-Community und ihren Projekten offen.

dhPLUS ist seit 1.1.2020 in Betrieb; der derzeit stattfindende Aufbau ist der Beitrag der PLUS zum KONDE-Projekt, aus deren Mitteln die Stelle eines *Repository Developers* finanziert wird.

dhPLUS war von Beginn an als ein Service der IT-Abteilung der PLUS (ITS) geplant, d. h. dass nach dem Releasedatum die für den Betrieb von dhPLUS zuständigen Stellen nicht aus Drittmitteln, sondern wie jedes andere Service des ITS aus Mitteln der Universität finanziert werden. Als Service des ITS nutzt dhPLUS auch die dort vorhandene IT-Infrastruktur, insbesondere die ORACLE-DB mit dem integrierten Archivsystem.

Der Aufbau der Plattform entspricht dem OAIS-Modell: Die Einspeisung der Daten in das Archivsystem erfolgt beim Ingest großer Datenmengen automatisiert, die Eingabe und die Bearbeitung einzelner Datensätze geschieht über Eingabemasken. Die Daten sind im Archivsystem in Digitalen Objekten abgelegt, die einen eindeutigen und persistenten Identifier (PID) haben und grundlegende Metadaten über das Digitale Objekt und über die darin enthaltenen Datensätze strukturiert und standardisiert bereitstellen. Jede Änderung an den Daten wird im Archivsystem versioniert. Der User-Zugang zu den Daten erfolgt in der Dissemination einerseits über eine Konvertierung der Daten in Standard-Formate wie .html, .pdf oder .xml, andererseits in der Bereitstellung von Applikationen, mit denen der User verschiedene Funktionen über den Daten ausführen kann. Angepasst an den generischen Aufbau von DH-Projekten wird nicht nur wie im OAIS-Modell die Datenschicht, sondern auch die Funktionsschicht archiviert. Für die LZA wichtig ist auch die dritte Art von Digitalen Objekten, die die (technischen und wissenschaftlichen) Dokumentationen eines Projektes beinhalten.

dhPLUS ist als Plattform auf Kontinuität und Stabilität hin ausgelegt. Zugleich sind die Informationstechnologien (IT) ein Bereich, der zu den kompetitivsten und innovativsten gehört. Es ergibt sich die Herausforderung, den Betrieb der Plattform auf Dauer zu gewährleisten und zugleich mit der rasanten technischen Entwicklung Schritt zu halten. Bei begrenzten finanziellen und personalen Ressourcen gelingt das nur mit einer weitgehenden Standardisierung der eingesetzten IT. Die diesbezügliche Strategie von dhPLUS ist, so weit wie möglich international anerkannte Standards zu übernehmen und nur dort, wo dies

nicht der Fall ist, auf weniger allgemeine Lösungen zurückzugreifen. Proprietäre Entwicklungen sind als Extremfall möglichst zu vermeiden.

Standardisierung wird auf verschiedenen Ebenen umgesetzt: In Bezug auf die Dateiformate werden alle Daten nach dem Ingest im Format .xml im Archivsystem abgelegt. Ausnahmen sind Projektdaten, die nicht in .xml, sondern nur in PDF/A gesichert werden können, und Pre-Ingest-Daten, die für die Rekonstruktion der Archivdaten nützlich sein können.

Eine weitere Vereinheitlichung betrifft die Auslegung von dhPLUS als *Linked Data*-Platform (LDP). Das bedeutet, dass alle im Archivsystem befindlichen Daten und Metadaten, die zur funktionellen Weiterverarbeitung verwendet werden, in RDF modelliert sind, wobei RDF in XML serialisiert ist. RDF ist ein Standard-Framework des *Semantic Web* und die Voraussetzung für LDP. Mit LDP ist es möglich, die Daten- und Metadatensätze der Digitalen Objekte via HTTP-*requests* zugänglich zu machen.

In RDF ist auch die weitere Content-Modellierung der Daten und Metadaten ausgeführt. Die Datensätze der bibliographischen, biographischen und topographischen Metadaten werden in eine CIDOC-Struktur eingefügt und mit Normdaten aus OWL-Ontolologien der GND und der WikiData gespeist. Die bibliographischen Metadaten werden in BIBFRAME modelliert, dem sich abzeichnenden Nachfolgestandard für das derzeit in den Bibliothekssystemen vorherrschende MARC21-XML-Format. Die MARC-Datensätze werden über Schnittstellen automatisiert aus den Bibliothekssystemen bezogen, nach BIBFRAME konvertiert und abgelegt. Über diese Schnittstellen geschieht auch ein automatisierter Abgleich der Datensätze. Gleiches gilt für die GND- und WikiData-Normdaten.

Zur Validierung der RDF-Daten wird SHACL auf dhPLUS implementiert. SHACL ist ein 2017 von der W3C herausgegebener Standard, der das Problem löst, dass es für RDF-Daten bisher kein mit dem XML-Schema vergleichbares Regelwerk gab und OWL dafür nur bedingt geeignet war. Mit SHACL können Datensätze auf Regelkonformität geprüft, neue RDF-*triples* erzeugt und SPARQL-*snippets* dort eingefügt werden, wo SHACL nicht ausreicht. Auf dhPLUS werden bestehende OWL-Ontologien nicht durch SHACL ersetzt, sondern in eine Kombination mit SHACL-Dateien integriert.

Die meisten DH-Projekte weisen Volltexte auf, die in TEI-XML modelliert sind. Auf dhPLUS werden diese archiviert und mit einer RDF-Repräsentation ergänzt. Zugleich werden alle TEI-Elemente und ihre RDF-Repräsentanten mit eindeutigen xml-IDs getaggt, so dass in den RDF-Statements die Informationen über die TEI und die ID enthalten sind. Annotationen werden über das *Web Annotation Vocabulary* mit den Statements verbunden. Das Resultat ist ein flexibles Annotationstool, mit dem ein *Stand off-Markup* realisiert werden kann, indem (theoretisch) beliebig viele Annotationsebenen mit den RDF-Repräsentanten der TEI-Elemente verbunden werden.

In der technischen Umsetzung ist dhPLUS als modulares System realisiert. Eine Folge davon ist, dass die ORACLE DB des ITS, an die die Plattform angebunden ist, kein systemimmanenter Teil von dhPLUS ist, sondern die Verbindung über ein Modul geschieht, das ersetzt werden kann, wenn die ORACLE DB durch eine andere Infrastruktur abgelöst wird. Gleiches gilt für die FEDORA-Architektur, die von den großen österreichischen Repositorien verwendet wird. dhPLUS ist kein FEDORA-Repositorium, es emuliert aber mit Hilfe von Modulen die FEDORA-Architektur, um die Zusammenarbeit und den Austausch mit den Repositorien der anderen österreichischen Universitäten zu erleichtern.

Kern des modularen Aufbaus von dhPLUS ist eine Microservice-Architektur. Prozesse des Ingest, der Archivierung und der Dissemination werden über Microservices abgewickelt. Jedes Microservice läuft in einem eigenen Docker-Container, der bei einer Fehlfunktion nur dieses eine Service, aber nicht die Plattform insgesamt zum Shutdown bringt. Ein Fehler-Monitoring-System kann die Fehlfunktion entdecken, das Service kontrolliert abschalten und als Ersatz ein anderes Service starten. Die Module sind skalierbar: Arbeitet ein Modul eine Aufgabe ab und tritt ein *request* nach der gleichen Aufgabe ein, kann ein weiteres Modul gestartet werden, das die gleiche Aufgabe übernimmt. Entspricht ein Microservice nicht mehr dem Stand der Technik oder muss aus einem anderen Grund ersetzt werden, kann das neue Modul im Testsystem geprüft und bei Erfolg in das Produktivsystem übernommen werden, ohne dass der Betrieb unterbrochen werden muss.

Der unterbrechungsfreie Betrieb bei gleichzeitiger Übernahme innovativer Entwicklungen ist ein Vorteil des modularen Aufbaus der Plattform und ein weiterer wichtiger Baustein für die Realisierung einer LZA von DH-Projekten. Vom Standpunkt des dhPLUS-Teams aus ist eine Langzeitarchivierung, die dadurch gewährleistet wird, dass es an einem Standort eine Plattform gibt, auf der Projekte dauerhaft serviciert werden, eine wichtige Voraussetzung, aber sie reicht nicht aus. Ein unwahrscheinlicher, aber möglicher Ausfall dieses Standortes und der Plattform würde den Verlust der Projekte und ihrer Daten bedeuten. Um einen solchen Fall zu verhindern, ist ein *deployment* der Plattform erforderlich, das in Zukunft etwa durch eine österreichweite Aufteilung der Plattform auf die verschiedenen Standorte erreicht werden könnte. Ein erster Schritt in diese Richtung könnte die Entwicklung einer Austauschontologie zwischen den Repositorien Österreichs sein, deren Konzeption im KONDE-Projekt bereits initiiert wurde.

Literatur:

- Hörmann, Richard; Schlager, Daniel: Saving Digital Humanities. In: digital humanities austria 2018. empowering researchers. Wien.

Software:

Apache Jena, Bootstrap, GND, Handle, iiif, latex, RDF, Solr, Wikidata

Verweise:

OAIS, Semantic Web, stand off markup, Datenmodell MHDBDB, DHA-Ontologie, Digitale Nachhaltigkeit, GAMS, Fedora

Themen:

Archivierung, Institutionen

Zitiervorschlag:

Hörmann, Richard. 2021. Disseminations-Services: dhPLUS. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.68