Fünf Thesen zum Thema Nachhaltigkeit

Die Sicherstellung der Verfügbarkeit von (Text-)Daten als Aufgabe von Langzeitarchivierung

Erfahrungsbericht aus einem nationalen Forschungsdateninfrastrukturprojekt

Johannes Stigler
Zentrum für Informationsmodellierung
Karl-Franzens-Universität Graz

Data Center for the Humanities der Universität zu Köln (DCH)
Cologne Center for eHumanities (CCeH)
in Kooperation mit der Landesinitiative NFDI der Digitalen Hochschule NRW
und der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste

17. September 2018

Versuch von Begriffsklärungen

„Nachhaltigkeit“

- Aus der Perspektive von Nutzer*innen
  Die Publikation von Digitalen Editionen ist ebenso einfach wie heute die eines PDF/A-Dokumentes in einem Repositorium und ist nicht an einen spezifischen Projektkontext gebunden
- Aus der Perspektive von Systemadministrator*innen
  Es gibt (a) eine Modellierungssprache für die Funktionalitäten von Digitalen Editionen und (b) eine standardisierte API.

„Kuratierung“

- Algorithmusgesteuerte Bestandserhaltung in einem sich in einem ständigen Wandel befindlichen IT-Ökosystem
- Erweiterung bestehender Projekte durch neue Funktionalitäten und neue Formen der Datenrepräsentation (Tools zur Visualisierung oder Analyse von (Text-)Daten, u.a.)

These I

Der Status quo zum Thema Nachhaltigkeit resultiert aus den Prämissen der Forschungsförderung

Infrastrukturprojekte im Bereich der Digitalen Geisteswissenschaften werden (in Österreich) nicht gefördert. Das Rad wird ständig neu erfunden. Dort wo sie gefördert werden, sind sie zu weit weg von den Bedarfen spezifischer Domänen[a].

[a] Data Management Pläne sind in Hinblick auf Nachhaltigkeit m.E. nur eine völlig unbefriedigende Zwischenlösung

These II

Digitale Langzeitarchive sind Publikationsinstanzen für Digitalen Editionen

Projektkuratierung erfolgt DANN nicht; muß also eingebettet sein in Prozesse der nachhaltigen Bestandserhaltung. Der Betrieb eines Langzeitarchivs erfordert eine technische und organisatorische Infrastruktur, sowie fachliches Knowhow rund um die Domäne aus der die zu verwaltenden Daten stammen[a].

[a] Danke für die ausführliche Klärung diese Tatbestandes im Grundsatzpapier der Arbeitsgruppe Datenzentren im Verband DHd

These III

Kuratierbarkeit ist ein notwendiges Strukturmerkmal nachhaltiger Repräsentationsform Digitaler Editionen

Die Möglichkeit zur Kuratierung ist konzeptionell bei der Entstehung von Daten und deren Repräsentationsweisen im Design von Editionsframeworks zu berücksichtigen.

These IV

Objektorientierung ist ein notwendiges Strukturmerkmal nachhaltiger Repräsentationsform Digitaler Editionen

Ein Framework, das geeignet ist, (Text-)Daten aus der Domäne der Digitalen Geisteswissenschaften aufzunehmen, ist kein System zur Verwaltung von Dateien, sondern erfordert eine objektorientierte Organisation seiner Inhalte gemäß dem Prinzip der Trennung von Inhalt und (Re-)Präsentation.

These V

Modellierung ist ein notwendiges Strukturmerkmal nachhaltiger Repräsentationsform Digitaler Editionen

Zur Sicherstellung von Nachhaltigkeit bedarf es standardisierter daten- (und nicht technologie)zentrierter Lösungsansätze in den Digitalen Geisteswissenschaften. Auch Repräsentationsschichten Digitaler Editionen sollten letztendlich modelliert und nicht programmiert werden.

KONDE – Kompetenznetzwerk „Digitale Edition“

Projektziele

- die Erarbeitung eines inhaltlichen und strategischen Konzepts zur Bündelung der Kompetenzen und den Aufbau einer nationalen digitalen Infrastruktur für Editionsprojekte
- die Definition von Workflows und Standards für die systematische und institutionenübergreifende Digitalisierung und Bereitstellung von Quellenmaterial
- die Erarbeitung einer Best-Practice Lösung für eine eng in LZA-Repositorien eingebundene Publikationsplattform für DE
- die Entwicklung von Konzepten für eine Übernahme des im Projekt entwickelten Referenzmodells für DE durch einschlägige Gedächtnisinstitutionen
- die (Weiter-)Entwicklung von Werkzeugen für die Verarbeitung des digitalen Materials mittels (semi-)automatischer Verarbeitungsschritte
- den Aufbau und die Etablierung einer einschlägigen Fortbildungsinfrastruktur zur Vermittlung der im Projekt erarbeiteten Standards

KONDE – Kompetenznetzwerk „Digitale Edition“

Zehn Partner

- Universität Klagenfurt (AAU)
- Adalbert-Stifter Institut des Landes Oberösterreich (ASI)
- Universität Innsbruck (IBK)
- Universität Graz (KFU)
- Kunstuniversität Graz (KUG)
- Österreichische Akademie der Wissenschaften (ÖAW)
- Österreichische Nationalbibliothek (ÖNB)
- Universität Salzburg (SBG)
- Technische Universität, Graz (TUG)
- Universität Wien (VIE)

KONDE – Kompetenznetzwerk „Digitale Edition“

Zehn Arbeitsgruppen

- Quellendigitalisierung und -beschreibung
- Transkription und Textauszeichnung
- Korpuslinguistische Analyse und Textmining
- Webpräsentationsformen und -interfaces
- Alternative Formen der Textcodierung
- Netzwerkanalyse und Datamining
- Archivierung
- Textkritik und -kommentar
- Strategiearbeitsgruppe
- Hybrid-Edition

KONDE – Kompetenznetzwerk „Digitale Edition“

Strategien zur Umsetzung der Projektziele

- Realisierung von digitalen Editionsvorhaben
- Arbeitsgruppen zu Schwerpunktthemen
- Workshops
- Weißbuch
- Öffentlichkeitsarbeit

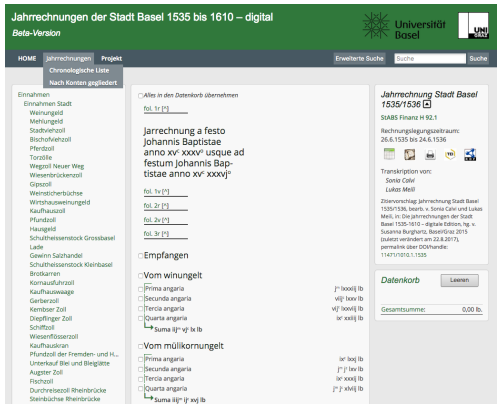
Objektorientierung: Ein Beispiel

- Jedes Objekt im digitalen Archiv hat einen komplexen Typ: Handschrift, museales Artefakt, Ontologie etc.)
- Heißt, jedem Objekt ist eine Klassendefinition zugeordnet …
  - mit einer vordefinierten (aber grundsätzlich erweiterbaren) Menge an Datenströmen
  - einer vordefinierten Menge an Workflows zur Auslieferung der Daten
  - einer vordefinierten Menge an Workflows, die beim Ingest oder der Änderung der Daten ausgelöst werden
  - und einem Regelwerk für die Erstellung von neuen Objekten (des jeweiligen Typs)

Objektorientierung: Ein Beispiel

Eine Handschrift aus der digitalen Edition der Basler Jahrrechnungen[a][b]

[a] http://gams.uni-graz.at/o:srbas.1535
[b] http://gams.uni-graz.at/context:srbas



Objektorientierung: Ein Beispiel

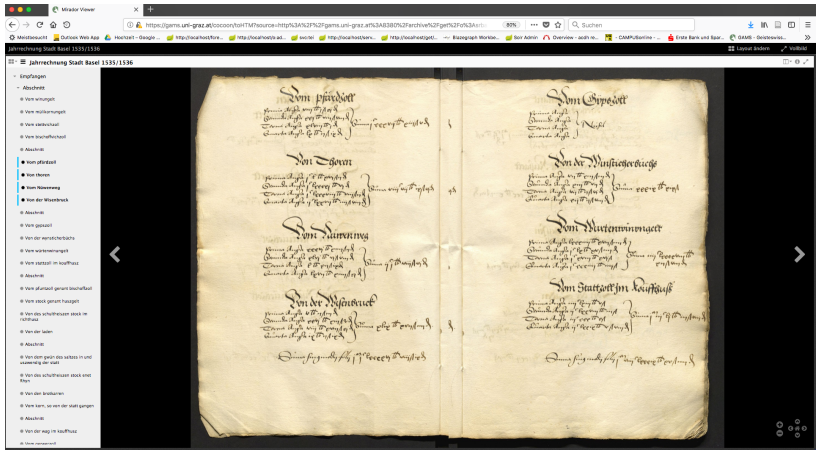
- Das Transkript der Handschrift ist gemäß den Richtlinien der TEI kodiert
- Beim Ingest des TEI-Dokumentes in das DLZA …
  - werden auch alle Faksimiles hochgeladen
  - wird das TEI-Dokument mit Metadaten aus externen Quellen (geonames.org, GND, SKOS Ontologien, etc.) angereichert
  - werden semantische Aussagen (RDF) aus dem TEI-Dokument extrahiert, im Objekt gespeichert und in eine semantische Datenbank übernommen
  - werden deskriptive Metadaten (DCMI, EDM u.a.) aus dem TEI-Dokument extrahiert und im Objekt gespeichert
  - wird – basierend auf den Strukturinformationen im TEI-Dokument – ein IIIF-Manifest erzeugt und im Objekt gespeichert

- On-the-fly, wenn ein Objektinhalt angefragt wird …
  - werden HTML, PDF, LaTeXPDF Repräsentationen des Objektes erzeugt und ausgeliefert
  - werden Faksimiles im DFG- oder Mirador-Viewer angezeigt
  - wird eine Datenmatrix aus dem Textkorpus nach einem Regelwerk erstellt und bereitgestellt

Objektorientierung: Ein Beispiel

Das TEI Objekt in der Anzeige des Mirador-Viewers[a]

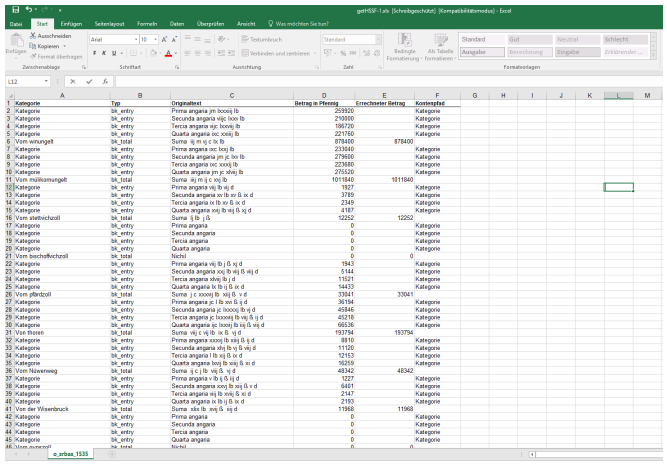
[a] http://gams.uni-graz.at/o:srbas.1535/sdef:IIIF/getMirador



Objektorientierung: Ein Beispiel

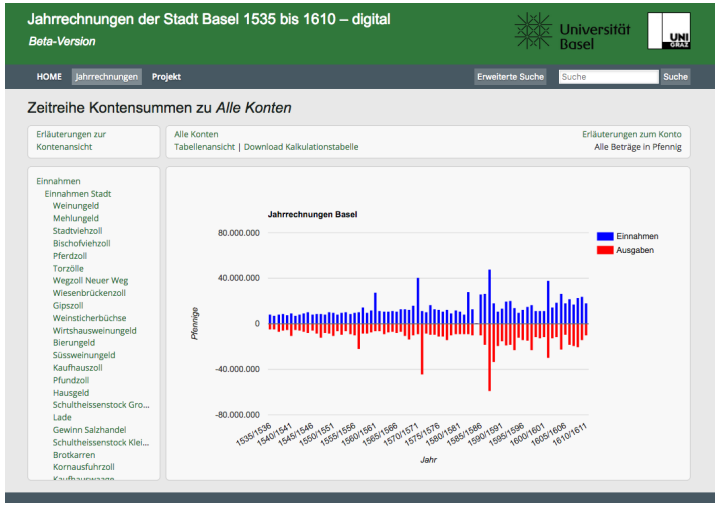
Die RDF-Daten des TEI Objektes als EXCEL Sheet[a]

[a] http://gams.uni-graz.at/o:srbas.1535/sdef:HSSF/get



Objektorientierung: Ein Beispiel

Ein auf Basis des gesamten TEI-Korpus generierter Summary Report



Objektorientierung: Ein Beispiel

…und all das aus dem TEI Quellcode[a]

[a] http://gams.uni-graz.at/o:srbas.1535/TEI_SOURCE

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:r="http://gams.uni-graz.at/rem/ns/1.0">
  <teiHeader xml:id="d2e4">
    <fileDesc xml:id="d2e6">
      <titleStmt xml:id="d2e8">
        <title xml:id="d2e10">Jahrrechnung Stadt Basel 1535/1536</title>
        <respStmt xml:id="d2e13">
          <resp xml:id="d2e15">Transkription von</resp>
          <persName xml:id="editor">Sonia Calvi</persName>
        </respStmt>
        <respStmt xml:id="d2e22">
          <resp xml:id="d2e24">Korrigiert von</resp>
          <persName xml:id="d2e27">Lukas Meili</persName>
        </respStmt>
        <principal xml:id="d2e31">Burghartz, Susanna (Universität Basel)</principal>
      </titleStmt>
      <publicationStmt>
        <distributor xml:id="d2e37">Zentrum für Informationsmodellierung - Austrian Centre for Digital
              Humanities, Universität Graz</distributor>
        <idno type="PID">o:srbas.1535</idno>
      </publicationStmt>
      <sourceDesc xml:id="d2e41">
        <msDesc xml:id="d2e43">
          <msIdentifier n="http://query.staatsarchiv.bs.ch/query/detail.aspx?ID=1137716" xml:id="d2e45">
            <repository xml:id="d2e47">StABS</repository>
            <collection xml:id="d2e50">Finanz H</collection>
            <idno xml:id="d2e53">92.1</idno>
          </msIdentifier>
          <msContents xml:id="d2e57">
            <p xml:id="d2e59">
              <origDate extent="52 Wochen" from="1535-06-26" to="1536-06-24" xml:id="d2e60"></origDate>
            </p>
          </msContents>
        </msDesc>
```


Kuratierbarkeit: Ein Beispiel

GAMS – Geisteswissenschaftliches Asset Managment System

- Eine Infrastruktur für unterschiedlichste Projekte (Editionen, Sprachkorpora, Epigraphische Sammlungen, Bildsammlungen u.a) mit je sehr spezifischen Funktionalitäten
- Mengengerüst des derzeitigen Datenpools: 90.000 Objekte mit rund 800.000 Datenströmen und 100.000.000 RDF-Triples
- Das älteste Objekt im Repositorium ist 14 Jahre alt

Kuratierbarkeit: Ein Beispiel

GAMS goes FEDORA 5.x

- Kompletttausch der Framework-Komponenten ohne Änderungen an den Einzelprojekten
- Vollständig Docker-basierte Systemarchitektur aller Services
- Implementierung der Content Model Logistik auf Basis von API-X OSGI Modulen
- Die Verlagerung vieler, derzeit clientseitig implementierter Workflows in ein Server-basiertes Doorkeeper-Webservice (fcgate) ermöglicht die Thin Client Entwicklung in beliebigen Programmiersprachen

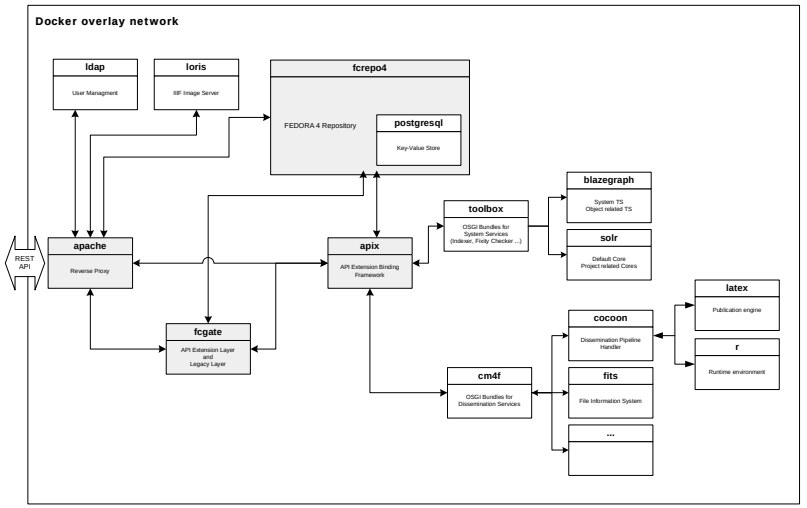
Kuratierbarkeit: Ein Beispiel

Summe der Erfahrungen aus der Testmigration

- Ausgangsbefund bei Testmigration mit Standardimage von FEDORA 4.7.5: Exponentiell ansteigende Ingest- und Requestzeiten
- Befund am Ende der Testphase: Stabile durchschnittliche Ingestzeit von rund 1.2 sec. pro Objekt und Requestzeiten im zweistelligen Millisekundenbereich.
- Schrauben an denen gedreht wurde: Tomcat-Konfiguration, PostgreSQL anstelle von Google Key-Value-Store, flachere Hierarchiestrukturen im Datenmodell bei gleichzeitiger Einführung von Zwischenschichten

Kuratierbarkeit: Ein Beispiel

Systemarchitektur unter FEDORA 5.x



Finis

Danke für Ihre Aufmerksamkeit!

http://gams.uni-graz.at
http://fedora-commons.org
http://apache.org
http://tei-c.org
https://iiif.io
http://lido-schema.org
http://geonames.org
u.a.