Ist die DARIAH-DE Forschungsinfrastruktur fit für Daten der realen Welt? Bericht über einen Anwendungsfall mit archäologischen Daten und seine ersten Ergebnisse

Romanello, Matteo
matteo.romanello@dainst.de
Deutsches Archäologisches Institut, Deutschland

Gradl, Tobias
tobias.gradl@uni-bamberg.de
Universität Bamberg, Deutschland

Hintergrund

Eine wesentliche Kritik an Forschungsinfrastrukturen behauptet:

This is the central paradox for big infrastructure design: the very wish to cater to everyone pushes the designers toward generalization, and thus necessarily away from delivering data models specific enough to be useful to anyone. (van Zundert 2012: 172)

Können generische Infrastrukturen und Datenmodelle für individuelle Forschungsfragen von Bedeutung sein? Und wenn ja, wie verhalten sich spezifische Forschungsdaten gegenüber den generischen Infrastrukturen? In diesem Beitrag diskutieren wir diese Frage im Hinblick auf die von DARIAH-DE [1] entwickelte Forschungsdateninfrastruktur und insbesondere das Data Modelling Environment (DME). Als Schema und Crosswalk Registry entstanden (Gradl et al. 2015), entwickelte sich das DME zu einem umfangreichen Werkzeug für die Modellierung, Verfeinerung, Bereinigung und Anreicherung von Daten. Die Beispieldaten, die für diesen Anwendungsfall herangezogen wurden, stammen aus einer Datenbank der aus dem Deutschen Archäologischen Institut geführten Grabung in Pergamon [2] und beschreiben etwa 100 keramische Grabungsfunde.

Für den Anwendungsfall wurde ein archäologischer Kontext gewählt, da relevante Forschungsdaten aufgrund ihrer Heterogenität eine besondere Herausforderung für Forschungsinfrastrukturen darstellen (Gradl, Henrich 2016a). Das wesentliche Ziel dieses Beitrags besteht darin, die Verwendbarkeit des DME auch im spezifischen Kontext von Pergamon-Daten anzudeuten. Eine Integration weiterer archäologischer Daten wie 2D-Bildern, 3D-Modellen und verschiedener Arten kontrollierter Vokabulare und geographischer Daten könnten für den gewählten Anwendungsfall Erkenntnisgewinne erreicht werden, die ggf. neue Fragen für die qualitative Forschung aufwerfen.

Durch eine kombinierte Visualisierung orts- und zeitbezogener Abhängigkeiten könnte man sich schnell einen ersten Überblick über die zeitliche und geographische Verteilung der Datensätze verschaffen. Wo liegt z. B. die höchste Dichte von auf die hellenistische Zeit datierten, keramischen Funden vor? Ein solcher visueller Überblick über die Grabungsdaten könnte den WissenschaftlerInnen auch erlauben, Diskrepanzen und Sonderfälle in der archäologischen Dokumentation einer Grabung zu erkennen.

Beschreibung des Anwendungsfalls

Die Verarbeitung der Daten wird unterstützt durch das DME und insbesondere dessen Fähigkeit, auf externe Ressourcen zuzugreifen. Zwei Schnittstellen zu Diensten des DAI wurden implementiert, damit zeit- und ortsbezogene Textangaben wie „Grobdatierung: hellenistisch-kaiserzeitlich“ oder „Provenienz: Pergamon“ mit den entsprechenden und in Zahlen ausgedrückten Werten kartiert werden können. Schließlich werden die angereicherten Daten mittels des DARIAH-DE Geo-Browsers visualisiert, um die zeitliche und geographische Verteilung der in den Datensätzen beschriebenen Objekte visuell abzubilden.

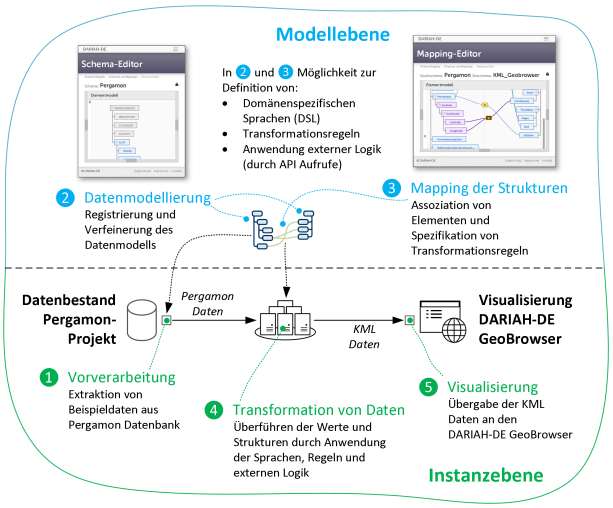

Abb. 1: Schematische Darstellung des Arbeitsablaufs.

Der Arbeitsablauf besteht aus fünf wesentlichen Schritten, die sich der *Modell-* oder *Instanzebene* zuweisen lassen.

- *Modellebene* : Datenmodelle und semantische Verbindungen zwischen diesen werden spezifiziert.
- *Instanzebene* : Während Aufgaben der Modellierung einmalig auszuführen sind, werden die Aufgaben der Instanzebene für jede Datei bzw. jede Aktualisierung der Daten durchlaufen.

Vorverarbeitung

Die archäologische Grabung in Pergamon wurde mittels der iDAI.Field Software dokumentiert. iDAI.Field ist ein modulares Dokumentationssystem für Feldforschungsprojekte, das am DAI entwickelt wurde und in ca. 50 verschiedenen Projekten eingesetzt wurde. [3] Die durch iDAI.Field gesammelten Daten werden in einer FileMaker-Datenbank gespeichert. Für eine Verarbeitung in der DARIAH-DE Infrastruktur wurde zunächst ein XML-Export aus der Datenbank ausgeführt.

Datenmodellierung

Um Pergamon-Daten in ein vom Geo-Browser unterstütztes Eingabeformat, wie die Keyhole Markup Language (KML) [4] umwandeln zu können, müssen die relevanten Datenmodelle im DME vorliegen bzw. definiert werden. Dies kann durch das Hochladen von XSD-Schemata initiiert werden. Einmal hinterlegte Modelle können nach deren Definition in weiteren Anwendungsfällen nachgenutzt werden.

Abbildung 2 veranschaulicht neben dem Elementmodell auch die Funktionalität zur Verarbeitung von Beispieldaten, mit deren Hilfe überprüft werden kann, ob Daten korrekt prozessiert werden.

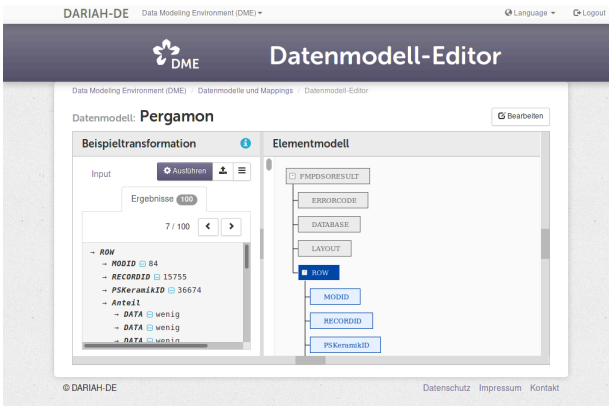

Abb. 2: Verarbeitung von Beispieldaten und Visualisierung der Datenstruktur.

Datenanreicherung

Die Funktionalität des DME stellt zwei wesentliche Methoden zur Datenmodellierung bereitgestellt (Gradl, Henrich 2016b):

- Inhaltliche Spezifikation von Daten durch die *Definition von domänenspezifischen Sprachen* (Parr 2012)
- Anwendung von Transformationsregeln. Neben bereits implementierter Funktionalität - z.B. zur automatischen Sprachverarbeitung - kann das DME flexibel durch Plugins erweitert werden, um neue Funktionen zur Verarbeitung von Daten einzubinden.

Im vorliegenden Anwendungsfall sind Daten in eine strukturierte und außerhalb der Pergamon-Datenbank interpretierbare Form zu bringen. Hierzu werden Daten u. A. durch die Nutzung des iDAI.Gazetteer [5] (Auflösung von Ortsbezeichnungen) und der iDAI.Chronontology [6] (Auflösung zeitlicher Angaben) angereichert.

iDAI.ChronOntology

Die ChronOntology API ermöglicht u. a. eine Freitextsuche. Beispielsweise ist es möglich nach Zeitangaben zu suchen, die den String „Kaiserzeitlich“ beinhalten [7] und die entsprechende Datierungen aufweisen können. So ist der Begriff „kaiserzeitlich“ mit „-27“ und „476“ als Beginn- und Enddatum verbunden.

Im Rahmen des DME wird das Modell der Pergamon-Daten dahingehend erweitert, dass unter dem in XML vorhandenen Element <Grobdatierung> zunächst Grammatik und Transformationsregel angelegt werden. Hierunter werden die zu produzierenden zusätzlichen Elemente modelliert: im konkreten Fall die strukturierte Antwort des iDAI.ChronOntology Dienstes. Abbildung 3 zeigt neben diesem erweiterten Elementmodell bereits das Ergebnis der Anwendung dieser Funktionalität auf die Beispieldaten.

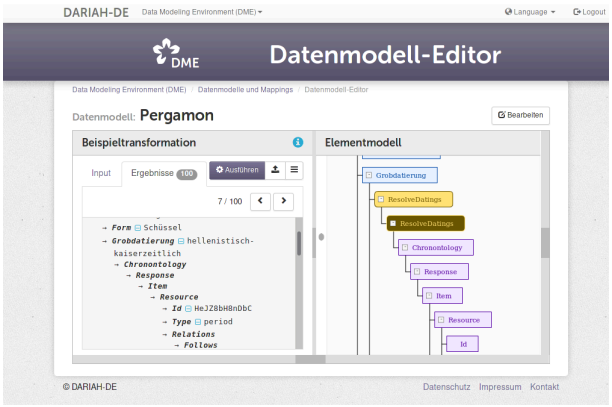

Abb. 3: Erweiterung des Datenmodells durch Zusatz der strukturierten Antwort des iDAI.ChronOntology Dienstes.

Für eine in Bezug auf die Pergamon-Daten optimierte Anfrage an den iDAI.ChronOntology Dienst wird die Semantik des Elements <Grobdatierung> expliziert. Die Grammatik in Abbildung 4 veranlasst die Zerlegung zusammengesetzter Datierungsangaben, um die vorliegende von-bis Semantik darzustellen (z. B. bei hellenistisch-kaiserzeitlich) und die einzelnen Anfrageterme zu extrahieren (Parr 2012).

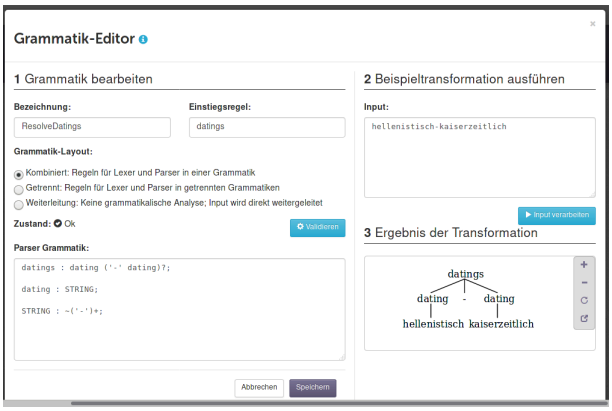

Abb. 4: Bearbeitung eines Elements des Datenmodells durch eine vom Benutzer editierbare Grammatik.

Durch die Adressierung der so gebildeten Terme in <dating> kann die anschließende Transformationsregel (vgl. Abbildung 5) auf eine verfeinerte Variante des zuvor unstrukturierten Inhalts zurückgreifen.

Die Ausführung der Chronontology API ist durch Anwendung von Funktionalität des umgesetzten DAI-Plugins möglich. Im vorliegenden Fall gestaltet sich das Kommando wie folgt:

Chronontology = dai.chronontology.query(@dating);

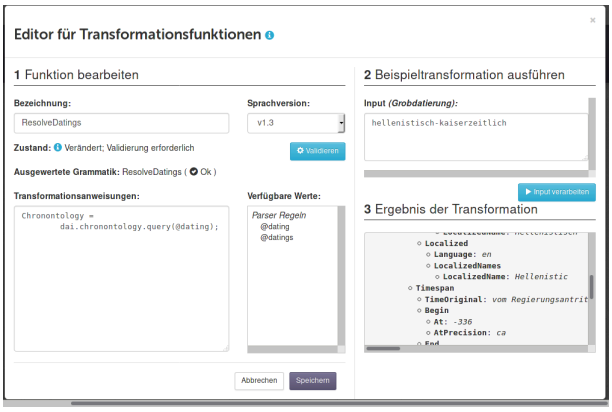

Abb. 5: Spezifikation einer Transformationsregel zur Abfrage des iDAI ChronOntology API.

Um aus der potenziellen Menge zurückgegebener Einträge ein Intervall zu berechnen, werden Kommandos aus dem math Funktionsraum verwendet:

BeginMin = math.min (@Response.Item.Resource.Timespan.Begin.At);
EndMax = math.max (@Response.Item.Resource.Timespan.End.At);

Hierdurch wird in den Daten exakt ein Zeitintervall hinterlegt, welches der gewünschten Semantik [frühester Begin, spätestes Ende] der Zeitangabe entspricht.

iDAI.Gazetteer

Vergleichbar mit der Chronontology API können auch Funktionen der Gazetteer API auf Daten angewandt werden. Im vorliegenden Beispiel wird der für eine Anfrage zurückgegebene, erste Treffer als wahrscheinlichste Koordinate verwendet und in den Daten berücksichtigt:

Location = dai.gazetteer.topcoord(@ResolveLocation);

Mapping der Datenstrukturen

Für die Transformation originärer Pergamon-Daten in das KML Format ist schließlich die Modellierung von Zusammenhängen der Datenmodelle erforderlich. Abbildung 6 zeigt die drei Mappings, die für den Anwendungsfall modelliert wurden.

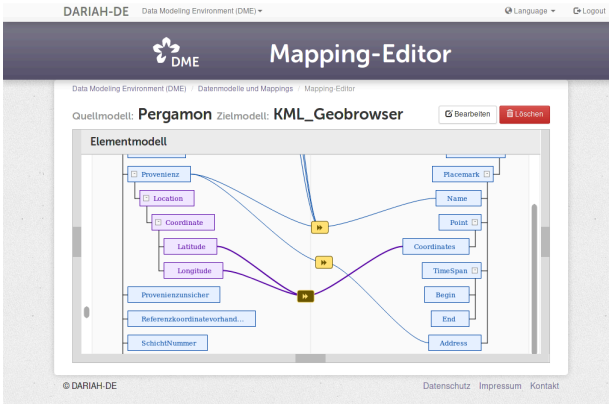

Abb. 6: Visualisierung der Mappings zwischen Quellmodell (Pergamon XML) und Zielmodell (KML) im DME.

Über einfache Wertkorrespondenzen, wie bei BeginMin (Pergamon) zu Begin (KML) hinaus können auch an dieser Stelle Transformationsregeln spezifiziert werden. Für die Übertragung der Koordinaten in das KML Schema wird so z. B. folgende Regel definiert:

[@Latitude != ""]
Coordinates = concat(@Latitude, ",", @Longitude)
[endif]

Koordinaten werden demnach nur angelegt, wenn @Latitude (für Daten im Quellschema) gesetzt ist. Zur Erzeugung eines Strings “Latitude, Longitude” wird die Konkatenationsanweisung verwendet.

Visualisierung der Mapping-Ergebnisse

Transformierte Daten können in verschiedenen Formaten heruntergeladen werden. Als KML Datei exportiert, können die 100 archäologischen Beispieldatensätze im GeoBrowser bereitgestellt und angezeigt werden (vgl. Abbildung 7).

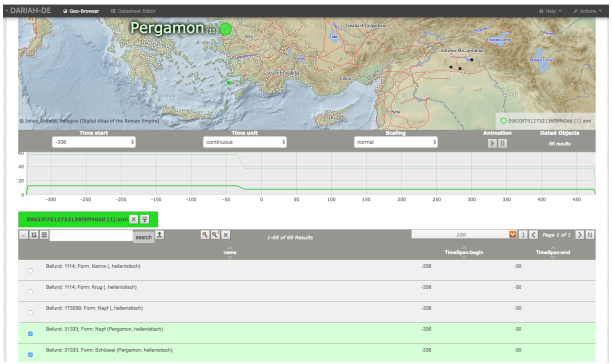

Abb. 7: Visualisierung der Mapping-Ergebnisse mittels des Geo-Browsers.

Nur 16 der 100 Datensätze haben Ortsangaben und können deshalb positioniert werden (Pergamon: 15, Knidos: 1), während fast alle eine Zeitangabe aufweisen. Die Möglichkeit, eine historische Karte (hier der *Barrington Atlas map of the Roman Empire*) auszuwählen, bietet einen zusätzlichen Nutzen, da sie eine bessere Kontextualisierung der Daten ermöglicht. Da der GeoBrowser derzeit keine XML-Namespaces unterstützt, müssen diese im Moment manuell aus den KML Daten entfernt werden.

Schlußdiskussion

Dieser Anwendungsfall basierte auf einer zu geringen Menge von Daten, als dass akute Mehrwerte erreichen werden könnten. Die Visualisierung von Daten aus mehreren Grabungsorten könnte dagegen die Einführung neuer Formen, Farben oder Keramiktypen in ort- und zeitbezogener Abhängigkeit darstellen und so zu der Generierung neuer Hypothesen führen. Das DME ist flexibel genug, um mit den heterogenen Daten der Archäologie umgehen zu können.

Indem das DME eine Modellierung von Verarbeitung von Daten spezifischer Anwendungsfälle ermöglicht, hat es das Potential das „OpenRefine für die digitalen Geisteswissenschaften“ zu werden: ein generisches Tool zur Modellierung, Verfeinerung, Bereinigung und Anreicherung von Forschungsdaten, das eine breite Vielfalt von Arbeitsabläufen unterstützen kann. [8]

Zugleich stellt sich aber auch die Frage, wer die typischen BenutzerInnen des DME sein können? Oder: ist es realistisch zu erwarten, dass GeisteswissenschaftlerInnen dieses Werkzeug ohne die Unterstützung von DH Spezialisten bedienen können? Tatsächlich scheint das DME eine gemeinsame Basis der Kollaboration und Kommunikation sein zu wollen, in der das Wissen von GeisteswissenschaftlerInnen mit der technischen Expertise von DH-Experten zusammengeführt werden. Hierdurch können Aufgaben, wie die des vorliegenden Anwendungsfalls erfüllt werden ohne sämtliche technische Problemstellungen von Grund auf neu lösen zu müssen. Durch die wachsende Zahl von bestehenden Quell-/ Zielmodelle, Transformationsregeln und API-Wrappers kann Wissen und Funktionalität nachgenutzt werden.

Fußnoten

1. *http://de.dariah.eu*
2. http://www.dainst.org/projekt/-/project-display/14186
3. https://www.dainst.org/forschung/forschung-digital/idai.welt/data/-/asset_publisher/Pt831IfwO8uH/content/idai-field
4. *https://wiki.de.dariah.eu/display/publicde/Geo-Browser+Dokumentation#Geo-BrowserDokumentation-Spezifikationenf%C3%BCrdieNutzung*
5. *https://gazetteer.dainst.org* . Vgl. auch Cuy et al. 2014.
6. http://chronontology.dainst.org/
7. Z.B. http://chronontology.dainst.org/data/period/?q=kaiserzeitlich
8. Für ein Beispiel der Benutzung von OpenRefine in den digitalen Geistenwissenschaften vgl. https://programminghistorian.org/lessons/cleaning-data-with-openrefine .

Bibliographie

van Zundert, J., (2012): “If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities”. *Historical Social Research*, 37(3), pp.165–186.

Cuy, Sebastian / Gerth, Philipp / Heiden, Maximilian / Kolbmann, Wibke / Schmidle, Wolfgang (2014): iDAI.gazetteer – ein Referenzsystem für altertumswissenschaftliche Ortsinformationen als Teil einer digitalen Forschungsinfrastruktur. In *Kölner und Bonner Archaeologica* 4, S. 203-212.

Gradl, Tobias / Henrich, Andreas (2016): Die DARIAH-DE-Föderationsarchitektur: Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen, Bibliothek Forschung und Praxis. Band 40, Heft 2, Seiten 222-228, ISSN (Online) 1865-7648, ISSN (Print) 0341-4183, DOI: https://doi.org/10.1515/bfp-2016-0027 , Juli 2016

Gradl, Tobias / Henrich, Andreas (2016): „Data Integration for the Arts and Humanities: A Language Theoretical Concept“. In: Fuhr, Norbert et al. (Hg.): Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice

of Digital Libraries, TPDL 2016, Hannover, Germany, September 5#9, 2016, Proceedings. Cham: Springer International Publishing, S. 281–293

Gradl, Tobias / Henrich, Andreas /Plutte, Christoph (2015): „Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Förderation von Kollektionen“. In: Baum, Constanze/Stäcker, Thomas (Hg.): Grenzen und Möglichkeiten der Digital HumanitiesZeitschrift für digitale Geisteswissenschaften. 2015, H. 1. URL: http://zfdg.de/sb001_020

Parr, Terence (2012): The definitive ANTLR 4 reference. 2. Aufl. Dallas, Raleigh: Pragmatic Bookshelf (= The pragmatic programmers)