APIS – Eine Linked Open Data basierte Datamining-Webapplikation für das Auswerten biographischer Daten

Schlögl, Matthias
Matthias.Schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Lejtovicz, Katalin
katalin.lejtovicz@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich

Einführung

Das ÖBL (Österreichisches Biographisches Lexikon 1815-1950) ist ein umfassendes Werk, das derzeit rund 18,000 Biographien von wichtigen historischen Persönlichkeiten aus der österreichisch-ungarischen Monarchie und der Ersten und Zweiten Republik Österreichs enthält. Während an dem Lexikon noch gearbeitet wird, erscheint es in gedruckter Form, und seit 2009 ist es auch online verfügbar.
APIS - Mapping historical networks: Building the new Austrian Prosopographical | Biographical Information System - ist ein interdisziplinäres Digital Humanities Projekt, das WissenschaftlerInnen aus unterschiedlichen Themenbereichen (Biografien, Geschichte, Geographie, Sozialwissenschaften, Informationstechnologie) verbesserten Zugriff (Suchabfragen, API etc.) auf die ÖBL-Daten erlauben wird. Dadurch wird es möglich sein innovative, interdisziplinäre Forschung auf der Grundlage dieser einzigartigen Ressource durchzuführen. Als erstes Beispiel für eine solche angewandte wissenschaftliche Forschung und als wichtiger Test der Brauchbarkeit und Eignung der entwickelten Lösung, wird bereits im APIS Projekt eine soziodemografische Analyse, die die Formen und Muster der Migration von gesellschaftlichen Eliten untersucht, umgesetzt.
In unserer Präsentation konzentrieren wir uns auf die zugrunde liegende technische Lösung, vor allem auf die dynamischen Aspekte - Workflow – und die Ergebnisse der verschiedenen angewandten Verfahren, um den aktuellen Stand der Umsetzung zu beschreiben.

Ansatz

ÖBL Daten stehen momentan in einem Ad-hoc-XML-Format zur Verfügung. Diese XMLs enthalten einige Fakten (Geburts- und Todesdaten, Orte, Berufsangaben usw.) in strukturierter Form, der Großteil der Information versteckt sich jedoch in dem unstrukturierten Haupttext der Biographie. Das Hauptziel des Projektes ist Informationen automatisch aus dem freien Text zu extrahieren, und sie in strukturierter Form zur Verfügung zu stellen. Um dieses Ziel zu erreichen, wird ein zweifacher Hybrid-Ansatz verfolgt, der einerseits automatische und manuelle Textverarbeitung kombiniert und andererseits erlaubt die erhobenen Daten in verschiedenen Formaten zu serialisieren. Letzteres beinhaltet nicht nur die Bereitstellung in verschiedenen Formaten (z.B. RDF/ JSON), sondern auch die Verwendung verschiedener Ontologien (z.B. CIDOC-CRM (Doerr 2003: 75-92), NDB (Historische Kommission bei der Bayerischen Akademie der Wissenschaften 1953)). Die extrahierten Entitäten sind mit mehreren semantischen Referenz Ressourcen wie zum Beispiel GND (Pfeifer 2012: 80-91), GeoNames oder DBpedia (Bizer 2009: 154-165) abgeglichen und mit URIs aus diesen versehen (Entity Linking). Dieser kombinierte Ansatz wurde gewählt, um die höchstmögliche Genauigkeit der Annotationen zu gewährleisten, und den manuellen Aufwand so gering wie möglich zu halten. Obwohl es bewährte Techniken und Methoden für die Verarbeitung natürlicher Sprache gibt, wird manuelle Arbeit (Korrektur) der Forscher, die mit den jeweiligen Wissenschaftsgebieten vertraut sind, nach wie vor erforderlich sein.

Datenmodell

Das Datenmodell besteht aus fünf Entitäten (Personen, Institutionen, Orte, Werke und Ereignisse) und einer Meta-Entität (Verweis auf den ursprünglichen Artikel). Es gibt Beziehungen zwischen allen Entitäten (z.B. Person - Institution, Person - Ereignis) und Beziehungen sind auch zwischen den gleichen Objekttypen möglich (z.B. Person -> Vater_von -> Person). Die Beziehungen können auch temporalisiert (Start- und Enddatum) und typisiert werden (Typen können je nach Bedarf angegeben werden). Das erlaubt uns praktisch alle möglichen Szenarien zu modellieren.
Der ursprüngliche Plan war, die Daten nach bestehenden, gut definierten Ontologien zu modellieren. In der Evaluierungsphase wurde uns aber klar, dass sehr viele verschiedene Ontologien existieren. Einige sind wie CIDOC-CRM Event basiert, andere verbinden Entitäten direkt. Wir haben uns deshalb entschlossen ein eigenes (internes) Datenmodell zu erstellen und so den technischen Aufwand für die Verarbeitung, Darstellung und Speicherung der Daten möglichst gering zu halten. Gleichzeitig werden wir aber dieses interne Datenmodell mit Hilfe schon existierender Ontologien (NDB, CIDOC-

CRM etc.) in verschiedenen Formen serialisieren und der Öffentlichkeit zur Verfügung stellen. Das stellt die möglichst einfache, nachhaltige Nutzung unserer Daten sicher.

Extraktion

Um strukturierte semantische Informationen aus den Biographien zu extrahieren, und die dadurch identifizierten Objekte zu Ressourcen wie GND, GeoNames zu verknüpfen verwenden wir automatische Tools. Die Ergebnisse werden von Experten verifiziert und ausgebessert um die Qualität der Daten zu gewährleisten, und um unser System durch manuelle Korrektur zu verbessern. Während die NLP-Tools eine schnelle Verarbeitung ermöglichen sind die Ergebnisse nicht zu 100% korrekt. Um die Genauigkeit zu verbessern, setzen wir mehrere Systeme, Quellen und Analysen ein. Für die automatische Extraktion haben wir mehrere Tools getestet und bewertet, wie z.B. Stanford NER (Finkel 2005: 363-370), GATE (Cunningham 2011), OpenNLP , Stanbol (Bachmann-Gmur 2013), basierend auf folgende Kriterien: 1) welche Sprachen unterstützt das System 2) Möglichkeit der Anpassung, 3) Entity Linking Fähigkeiten, 4) Output Format und 5) die Verfügbarkeit und Qualität der API. Apache Stanbol hat sich als das am besten geeignete Werkzeug für unsere Zwecke gezeigt. Stanbol ermöglicht die Verknüpfung von Entitäten wie Personen, Institutionen zu Referenzressourcen (Normdateien, Ontologien). Wir haben die Biographien mit GND und GeoNames abgeglichen, und planen weitere LOD Ressourcen hinzuzufügen. Durch die Verknüpfung von oben benannten Entitäten zu den semantischen Ressourcen können wir viele zusätzliche Informationen (z.B. Alternative Namen, Titel von Werken usw.) zu unseren Daten hinzufügen, und so Inhalte mit fehlenden Informationen bereichern.

Anwendung

Um den manuellen Arbeitsaufwand (Korrektur der Daten etc.) zu minimieren haben wir eine effiziente und einfache Weboberfläche geschaffen, die es den ForscherInnen erlaubt mit den Daten zu interagieren. Im Sinne einer nachhaltigen Nutzung und einfacher weiteren Betreuung des so entstandenen Tools haben wir uns entschlossen auf erprobte Web-Technologien zu setzen (Django /MySQL). Die Web-Anwendung ist in Django, einem Python-basierten Web-Entwicklungs-Framework, implementiert. Django ist nicht nur ein ausgereiftes und verbreitetes Tool (Websites wie Disqus, Pinterest und die Washington Times nutzen es), sondern bietet auch die Möglichkeit die volle Bandbreite der verschiedenen Python Bibliotheken nativ im Code zu verwenden (NLTK , scikit-learn , NumPy etc.) .
Die Web-Anwendung stellt die Daten der einzelnen Biographien strukturiert in drei Teilen dar: primäre minimale Informationen, Haupttext mit markierten Anmerkungen und die Listen von Orten, Institutionen und Personen, die mit dem Biographierten in Zusammenhang stehen. Die Anwendung bietet auch Funktionen für die Navigation: dropdown Listen sowie einfache Volltextsuche.
Eine weitere wichtige Funktion der Anwendung ist die Möglichkeit, den Text manuell mit Annotationen zu versehen. Dieses Feature erlaubt sowohl die Korrektur von automatischen Annotationen, als auch das Hinzufügen von neuen Annotationen. Die Kuratoren können die Entitäten mit der Maus auswählen oder im Kontextmenü identifizieren.

Derzeit liegt der Schwerpunkt auf der Darstellung von Orten. Dementsprechend wird die Anwendung mit eingebetteten Karten ausgestattet, an denen identifizierte geographische Orte visualisiert werden können. In der nächsten Phase des Projekts wird eine interaktive Visualisierung entwickelt, um das Verständnis der Daten und die Navigation im Datenbestand zu erleichtern.

Arbeitsablauf

Das System unterstützt zwei Workflows: im ersten Schritt schickt die Anwendung (das Extrakt-Modul) die Biographien im Batch-Modus zu einem Extraktionsservice (lokale Stanbol Instanz), welches die Abfragen an externe Services und/oder lokale Indizes weiterleitet und die gematchten Entitäten in einer Liste in JSON-LD Format zurückgibt. Diese Entitäten werden von dem Extrakt-Modul analysiert und in der Datenbank abgespeichert. Danach werden sie in der Web-Anwendung dargestellt und können von den ForscherInnen überprüft und korrigiert werden.
Im zweiten Schritt wird der Workflow vom Benutzer gestartet: Der menschliche Annotator markiert einen String und identifiziert ihn als Ort, die Anwendung schickt den ausgewählten String zur Stanbol Instanz, die die verfügbaren Ressourcen abfragt und mögliche Kandidaten zurückgibt. Diese Treffer werden dem/der ForscherIn in Form eines Autocomplete Feldes angezeigt.

Schlussfolgerung

Während wir uns in unserem Abstrakt auf die technische Umsetzung konzentriert haben, ist es wichtig im Auge zu behalten, dass das System nur eine Voraussetzung ist die eigentliche Forschungsfragen beantworten zu können. Alle im Projekt generierten Daten sowie die entwickelte Forschungsumgebung wird der Öffentlichkeit zugänglich gemacht (eine erste Version der Forschungsumgebung wird Ende September in unserem Github Account zugänglich gemacht). Wie schon weiter oben angesprochen versuchen wir die Nachhaltigkeit unserer Lösung auf mehrfache Weise zu erreichen. Zum einen verwenden wir gut etablierte Web-Technologien und ermöglichen somit vielen Entwicklern weltweit unseren Code zu warten und/oder

weiter zu entwickeln. Zum anderen verbinden wir unsere Daten mit der LOD-Cloud und serialisieren sie mit Hilfe verschiedener weit verbreiteter Ontologien in den gängigsten Formaten und stellen so sicher, dass andere Projekte unsere Daten mit äußerst kleinem Aufwand direkt in ihre Projekte einbetten können.

Fußnoten

1. http://www.geonames.org/
2. https://opennlp.apache.org/
3. http://linkeddata.org/
4. https://www.djangoproject.com/
5. http://www.nltk.org/
6. http://scikit-learn.org/stable/
7. http://www.numpy.org/

Bibliographie

APIS: Mapping historical networks: Building the new Austrian Prosopographical | Biographical Information System (APIS) http://www.oeaw.ac.at/acdh/en/apis

Bachmann-Gmur, Reto (2013): *Instant Apache Stanbol* (1st ed.). Packt Publishing. ISBN 1783281235.

Bizer, Christian / Lehmann, Jens / Kobilarov, Georgi / Auer, Soren / Becker, Christian / Cyganiak, Richard / Hellmann, Sebastian (2009): „DBpedia - A crystallization point for the Web of Data“, in: *Journal of Web Semantics* 7 (3): 154–165.

Cunningham, Hamish / Maynard, Diana / Bontcheva, Kalina (2011): *Text Processing with GATE* (Version 6). University of Sheffield Department of Computer Science. ISBN 0956599311.

Doerr, Martin (2003): „The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata“, in: *AI Magazine* 24 (3): 75–92.

Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher (2005): „Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling“, in: *Proceedings of ACL-2005* 363–370.

Historische Kommission bei der Bayerischen Akademie der Wissenschaften (seit 1953): *Neue deutsche Biographie*, Berlin: Duncker & Humblot. ISBN 3-428-00181-8

ÖBL - Österreichisches Biographisches Lexikon/ Austrian Biographical Lexicon (1815-1950) Online-Edition und Österreischiches Biographisches Lexikon ab 1815 (2. Überarbeitete Auflage - online). Verlag der Österreichischen Akademie der Wissenschaften. Wien. http://www.biographien.ac.at/oebl [letzter Zugriff 26. August 2016]

Pfeifer, Barbara (2012): „Vom Projekt zum Einsatz. Die gemeinsame Normdatei (GND)“, in: Brintzinger, Klaus-Rainer (ed.): *Bibliotheken: Tore zur Welt des Wissens*. 101. Deutscher Bibliothekartag in Hamburg 2012, Olms, Hildesheim u.a. 2013: 80–91.