Linguistic texts – enhancing PubMan

Persönliche Dokumentsammlungen am Beispiel der Erschließung linguistischer Literatur aus dem Netz für die wissenschaftliche Arbeit

Institute: MPI-EVA

Responsible for the proposal: Prof. Dr. Bernard Comrie (Direktor Abteilung Linguistik, MPI-EVA)
Malte Dreyer (Leiter Abteilung Forschung & Entwicklung, MPDL)

Requested Budget (total): **192.500 Euro**

Requested Budget (annual):
1. Year: 18.667 Euro
2. Year: 112.000 Euro
3. Year: 61.833 Euro

Zusammenfassung

Am MPI für Evolutionäre Anthropologie (MPI-EVA) in Leipzig ist eine Sammlung von wissenschaftlichen Arbeiten entstanden, vorwiegend aus dem Fachgebiet Linguistik, die in elektronischer Form frei zugänglich im Internet aufgefunden wurden. Diese sollen für die wissenschaftliche Nutzung erhalten und weiter erschlossen werden unter Erweiterung und Einsatz der von der MPDL im Rahmen von eSciDoc entwickelten oder lizenzierten Instrumente *PubMan*, *Lucene* u.ä.
Persönliche Sammlungen von digitalen Dokumenten anzulegen und Kollegen darauf Zugriff gewähren zu können, ist eine Anforderung, die viele Wissenschaftler der MPG haben.

Projektziele sind daher zum Einen die Erschließung und Rechteklärung der Dokumente des MPI-EVA und zum Anderen die Entwicklung der nachnutzbaren Lösung für persönliche Dokumentsammlungen, auf die auch Dritten Zugriff gewährt werden kann.

Die in diesem Projekt entwickelten Arbeitsverfahren und Methoden wie die Referenzierung einzelner Textstellen, das Personalisieren von Dokumentsammlungen, das Gewähren von Zugriffsrechten für Dritte oder der Workflow zur Klärung und Dokumentation der Urheberrechte sind sowohl für andere Institute als auch für die MPDL von Nutzen für weitere Projekte.

Detaillierte Projektbeschreibung

Wissenschaftlicher Hintergrund

Sammlung ephemerer wissenschaftlicher Dokumente

Konkretes wissenschaftliches Ziel dieses Projektes ist es, linguistische Arbeiten, die im Netz frei zugänglich sind, einzusammeln und für die Forschung am Institut nutzbar zu machen. Da sich die Abteilung Linguistik mit Sprachen der ganzen Welt beschäftigt, sind v. a. Beschreibungen seltener, entlegener Sprachen interessant. Diese Werke sind oft frei zugänglich, da sie keinen kommerziellen Wert haben und nicht in einem Verlag veröffentlicht werden. Sie stehen häufig auf den Webseiten des Autors, einer Institution, einer Fachgesellschaft oder eines regionalen Kulturverbandes. Meist handelt es sich um ältere Arbeiten (5 Jahre oder mehr). In der Regel liegen sie im PDF-Format vor.
Gemeinsames Kennzeichen all dieser Werke ist ihre Flüchtigkeit. Es handelt sich in diesen Fällen nicht um Veröffentlichungen etablierter Verlage, Repositorien o.ä., sondern um Websites, die leicht und unbemerkt wieder verschwinden, umziehen oder geändert werden. Damit sind diese Arbeiten häufig wieder verloren. Die Auswahl der Werke erfolgt durch Wissenschaftler der Abteilung Linguistik.

Es erfolgt keine systematische oder automatisierte Suche mit Robotern o. ä. sondern die Suche wird gesteuert vom fachlichen Interesse und vom Zufalls- bzw. Schneeballempfehlungsprinzip.

Wissenschaftlicher Nutzen
Richtig erschlossen für die wissenschaftliche Nutzung verspricht die Sammlung sprachenbezogener Dokumente neue Forschungsmöglichkeiten, die so in der vergleichenden Linguistik bis jetzt nicht möglich waren. Ein wichtiger Teil der linguistischen Forschung am MPI-EVA besteht aus die Auswertung bestehender Beschreibungen der Sprachen der Welt, die normalerweise nur in klassischer gedruckter Form vorhanden sind. Eine gezielte Volltextsuche würde die Forschungsarbeit hier sehr erleichtern und verbessern. Wie die Erfahrung mit Webseiten zeigt, erleichtert die dauerhafte Referenzierbarkeit (über Links) von Textstellen die wissenschaftliche Arbeit mit digitalen Dokumenten.

Persönliche Dokument- und Referenzsammlungen
Die Möglichkeit, persönliche Sammlungen solch digitaler Dokumente und Links auf Textstellen in PDF- oder Bild-Dokumenten wissenschaftlich zu nutzen, wäre nicht nur für die Forschung am MPI-EVA, sondern auch in anderen Instituten von großem Vorteil.

Volltextrecherche
Einer der Wünsche, die der schon bestehenden Sammlung zu Grunde liegen, ist es, Volltext-Recherchen über alle vorhandenen Werke durchführen zu können. Derzeit können nur einzelne Werke durchsucht werden (durch OCR-Bearbeitung), aber nicht mehrere Werke parallel. Eine solche Volltext-Suche sollte auch auf die schon im Bibliothekskatalog vorhandenen Information zurückgreifen können (wie, z.B. Sprach-Codes).

Granulare Referenzierung
Eine Online-Plattform zur Durchsuchung aller gesammelten Werke wird noch wertvoller, wenn auch die Möglichkeit persönlicher Link-Sammlungen auf Textstellen innerhalb der gesammelten Dokumente (z.B. im XML- oder PDF-Format) gegeben ist. Deshalb hat dieses Projekt zum Ziele, solche Sammlungen über einen Dienst in PubMan bereitzustellen. Diese Erweiterung soll das Sammeln von eindeutigen und stabilen Referenzen (Links) auf Textstellen, Dokumente und Literatursammlungen erlauben. Referenzen auf einzelne gefundene Stellen können dann vom Forscher für spätere Auswertungen aufbewahrt, und mit Wissenschaftspartnern ausgetauscht werden. Da diese Referenzen stabil sein müssen (Persistent Identifier), eignen sie sich auch für die Ablage in Datenbanken und sogar für wissenschaftliche Literaturverweise. Ein solcher Dienst wird sich auch für andere Institute als nützlich erweisen.

Flexibilisierung der Zugangsbeschränkungen
Der Ansatz des Sammelns linguistischer Texte aus dem Netz erfordert aber eine noch genauere Rechteklärung und –dokumentation. Da das MPI-EVA im Moment nicht die Möglichkeit hat, alle Informationen zu organisieren, wird momentan nur eine teilweise Dokumentation der Rechtesituation festgehalten. Mit einer genauen Rechteklärung sollte es ermöglicht werden, die Vergabe der Zugangsrechte flexibler zu gestalten. Eine flexible Zugangsbeschränkung kann z.B. auch genutzt werden, um lokal Bände zu digitalisieren und lokal online zur Verfügung zu stellen, wenn das fachlich in einem Forschungsprojekt sinnvoll erscheint.

Die Entwicklung des Arbeitsverfahrens für diese Klärung ist über den speziellen Fall an diesem MPI hinaus für alle MPIs relevant, da sie auch auf Literatur anderer Fachgebiete übertragbar ist und daher von jedem anderen MPI nachgenutzt werden kann.

Internationale Kooperation
Ein flexiblerer Zugang zu den Werken wird auch benötigt, um verschiedene Kooperationen mit Universitäten in Ländern der sog. zweiten oder dritten Welt weiter auszubauen. Wir sind sehr interessiert an Magister- und Doktorarbeiten aus Ländern, in denen viele Studenten Arbeiten über ihre eigene Sprache schreiben, diese Arbeiten aber nie im wissenschaftlichen Mainstream ankommen. Projekte mit Universitäten in Kamerun, Nigeria, Äthiopien, Laos, Thailand und Brasilien sind im Gang oder in Vorbereitung, deren Ergebnisse schließlich auch in PubMan eingepflegt werden sollen. Außerdem ist eine flexible Rechtevergabe notwendig, um gemeinsam mit externen Wissenschaftlern mit persönlichen Dokumentsammlungen in PubMan arbeiten zu können.

Personelle Erfordernisse

Für die in diesem Zusammenhang am Institut für evolutionäre Anthropologie anfallenden Tätigkeiten soll eine bibliothekarische Fachkraft eingestellt werden. Wir beantragen daher die Finanzierung einer befristeten Stelle für eine(n) Diplombibliothekar(in) am MPI für evolutionäre Anthropologie durch die MPDL. Daneben ist es notwendig, PubMan so zu erweitern, dass den wissenschaftlichen Anforderungen des Institutes an eSciDoc entsprochen werden kann. Dafür ist auf Seiten der MPDL eine befristete Entwicklerstelle (wissenschaftlicher Entwickler) nötig.

Vorgesehene Aufgaben für Diplombibliothekar/in (Track A)

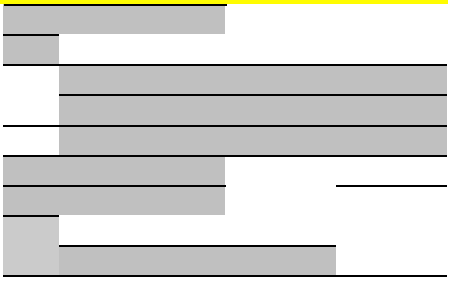
- Prüfung der bibliographischen Einbindung der eingespielten Daten in PubMan, Definition benötigter Anpassungen, usw.;
- Prüfung der Funktionalitäten der Volltextsuche mit *Lucene* als integralem Bestandteil von eSciDoc;
- Spezifikation zusätzlicher Elemente zur inhaltlichen Erschließung für das Metadatenprofil und die entsprechende PubMan-Eingabemaske, z.B. Verknüpfung mit Sprach-Codes (ISO 639-3/639-4), geographischen Informationen, Inhaltsverzeichnisse, Referenzierung auf Zitatstellen u.ä.;
- Qualitätskontrolle der eingepflegten PDF-Dokumente
- Einbinden der in PubMan eingepflegten Dokumente in den lokalen Bibliothekskatalog durch die von PubMan vergebenen Persistent Identifiers für die weitere wissenschaftliche Benutzung (z.B. Literaturverweise, Datenbankeinbindung);
- Ausbau der Sammlung durch Einpflegen weiterer Dokumente in Zusammenarbeit mit den Wissenschaftlern vor Ort;
- Aufbau eines institutsinternen Workflows zur Klärung der Urheberrechte, d.h. Ermittlung und Dokumentation der vorliegenden Rechte für die einzelnen Dokumente, dementsprechend gestaffelte Freischaltung der Dokumente (institutsintern oder weltweit zugänglich) sowie bei Bedarf Einholung eines einfachen Nutzungsrechts zur Online-Publikation von den Autoren;
- Dokumentation des Vorgehens zur Ermöglichung der Nachnutzung an anderen Instituten.

Vorgesehene Aufgaben für wissenschaftl. Entwickler/in (Track B):

- Erweiterung der Eingabemaske von PubMan um Eingabefelder für Sprach-Codes, und geographische Informationen
- Konzeption und Implementierung eines Services zur Erstellung und Verwaltung persönlicher Sammlungen von digitalen Dokumenten und von Referenzen auf und zwischen Textstellen in XML- und PDF-Dokumenten und anderen Dokumentsammlungen/Inhaltsverzeichnissen in eSciDoc.

Zeitplanung

Track	Paket	Jahr 1	Jahr 2
A	Mitarbeit an der PubMan-Erweiterung		
	Spezifikation zusätzlicher Eingabe-Elemente		
	Qualitätskontrolle		
	Verknüpfen der in PubMan bef. Dokumente mit Bibliothekskatalog		
	Ausbau der Sammlung		
	Aufbau eines Workflows von Sammlung bis zur Veröffentlichung		
	Dokumentation des Workflows		
B	Erweiterung der PubMan Eingabemaske		
	Service z. Verwaltung v. Referenzsammlung		



Beantragte Ressourcen: