Beitrag zum GEO-LEO Artikel in »Fachportale in Bibliotheken«

## Technik

Die Umsetzung des GEO-LEO Portals beruht auf mehreren Technologien, die im Folgenden vorgestellt werden:

1. Metasuch-Service Pazpar2
2. OAI Harvesting
3. Content-Management-System TYPO3 mit Erweiterungen
4. DSpace Dokumentenserver

Die Auswahl dieser technischen Lösungen erfolgte mit Blick auf die Wartbarkeit und das Vorhandensein von Know-How zu der jeweiligen Software. Sämtliche Softwarekomponenten sind an der SUB Göttingen auch zu anderen Zwecken (andere Fachportalte und Dokumentenserver) im Einsatz und unter freien Lizenzen verfügbar.

### Pazpar2

Pazpar2 ist eine Open Source Metasuchsoftware von Index Data [\*](http://www.indexdata.com/pazpar2). Sie führt Metasuchen über die Z39.50 und SRU Protokolle sowie Solr Indexe durch.

Pazpar2 ist ein reiner Datendienst: Es stellt die Antworten auf Suchanfragen als Daten zur Verfügung. Die Nutzeroberfläche zur Darstellung der Ergebnisse ist ein separates Modul. Hierdurch kann der Metasuchdienst unabhängig von der Technik und dem Design der Website, in der die Daten erscheinen, betrieben werden. Durch eine Schnittstelle über das Web-Protokoll http läßt sich Pazpar2 gut in Webseiten einbinden. Die Software ermöglicht eine dynamische Anzeige mit schnellem Erscheinen der ersten Treffer, indem sie gleich mit dem Eintreffen der ersten Ergebnisse der Metasuche beginnt, Treffer auszuliefern. So läßt sich eine dynamische Benutzeroberfläche realisieren.

Die Arbeit von Pazpar2 besteht aus mehreren Phasen:

1. Umwandeln der Abfrage des Nutzers in die Abfragesprachen der Zieldatenbanken
2. Stellen der Anfragen an die Zieldatenbanken mit dem jeweils passenden Protokoll
3. Einsammeln der Ergebnisse von den Zieldatenbanken
4. Normalisieren der Ergebnisse
5. Ausliefern der aktuellen Ergebnismenge

#### Abfragesprachen

Pazpar2 akzeptiert strukturierte Abfragen in der *Common Command Language* (CCL, ISO 8777). In dieser Sprache lassen sich Boolesche Abfragen auf verschiedenen Feldern ausdrücken. Beispiel: Die Abfrage „Grönland and person="Wegener, A\*"“ für eine Suche in allen Feldern nach „Grönland“ und im Feld „person“ rechtstrunkiert nach der Phrase „Wegener, A“.

Die hierbei verfügbaren Feldnamen werden in der Pazpar2 Konfiguration eingestellt und auf die Feldnamen bzw -nummern der jeweiligen Zieldatenbank abgebildet. Die boolesche Struktur der Abfrage wird in die für die Zieldatenbank nötige Abfragesprache übersetzt. Die Beispielabfrage wird für eine SRU Abfrage im CCL Format belassen, die Feldnamen aber angepaßt: „Grönland AND pica.per:"Wegener, A\*"“; für eine Z39.50 Abfrage wird sie in die Prefix Query Format (PQF) Abfrage „@and @attr 1=1016 Grönland @attr 1=1003 @attr 5=1 "Wegener, A"“ mit den entsprechenden Bib-1 Attributen konvertiert; und für Solr Indexe wird die Solr Abfrage „Grönland AND pers:"Wegener, A\*"“ mit den für diese Schnittstelle konfigurierten Indexnamen verwendet.

#### Abfrageprotokolle

Pazpar2 kann Abfragen über die Z39.50 und SRU Protokolle sowie über die http Schnittstelle von Solr ausführen. Die nötigen Details wie Protokollversionen und Zeichencodierungen sind hierbei wenn nötig konfigurierbar.

#### Einsammeln der Ergebnisse

Pazpar2 sammelt die Suchergebnisse über das jeweilige Protokoll in kleinen Mengen à 20 Datensätze ein. Hierdurch sind die ersten Ergebnisse fast sofort verfügbar und können schnell an die Endnutzer weitergeleitet werden.

Die Ergebnisse müssen als XML oder im ISO 2709 MARC Format vorliegen, um von Pazpar2 bearbeitet zu werden. Unstrukturierten Text (SUTRS) oder andere Binärformate wie MAB2 liest die Software nicht.

#### Normalisierung

Der wichtigste Schritt der Datenverarbeitung durch Pazpar2 ist die Normalisierung. Hierzu definiert die Konfiguration die Felder eines internen Datenmodells, in die die Felder der geladenen Daten abgebildet werden müssen. Für Daten im MARC Format liefert die Pazpar2 Installation mit der „tmarc.xsl“ Datei eine gute Standardkonversion mit, die viele Felder extrahiert. So läßt sich mit wenig Aufwand eine gute Einbindung von MARC Daten erreichen.

Um reichhaltigere Daten für die Anzeige zur Verfügung zu stellen, haben wir diese Konversion erweitert, um spezielle Felder aus dem Bereich Geographie (Koordinaten von Landkarten) oder dem deutschen Bibliothekswesen (ZDB-Nummer) zu berücksichtigen. Eine Normalisierung der verfügbaren Medientypen auf eine überschaubare Anzahl und der Sprachcodes auf den ISO 639-2/B Standard ermöglicht hilfreiche Facetten zum Filtern der Ergebnisliste.

Die Normalisierung funktioniert als XML / XSL Workflow: Daten werden als XML eingelesen, beziehungsweise von ISO 2709 MARC Format automatisch in XML umgewandelt. Für jede Zieldatenbank wird eine Folge von XSL Dateien festgelegt, die auf jeden Datensatz angewandt wird. So können wiederkehrende Konversionsschritte als einzelne XSL Dateien abgelegt und nach Bedarf kombiniert werden.

Einige Felder des internen Datenmodells können zur Nutzung für die Deduplizierung markiert werden: Pazpar2 fügt dann alle Datensätze zusammen, bei denen der Inhalt dieser Felder übereinstimmt. So können – eine einigermaßen homogene Katalogisierung vorausgesetzt – mehrfache Treffer für denselben Titel oder auch verschiedene Erscheinungsformen (z.B. elektronisch und gedruckt) oder Auflagen desselben Buches zu einem Treffer in der Ergebnisliste zusammengeführt werden. Die Detailinformation aus den verschiedenen Datenquellen bleiben hierbei erhalten und können in die Darstellung einfließen.

#### Ausliefern der Ergebnisse

Die Kommunikation mit Pazpar2 geschieht über das http Protokoll. Sobald eine Suche gestartet ist, ist ein Abruf der bereits geladenen Treffer möglich. Damit können fast sofort erste Ergebnisse der Suche angezeigt werden.

#### Konfiguration

Die Pazpar2 Konfiguration liegt in XML Dateien. In ihnen werden für den Pazpar2 Server verschiedene „Services“ eingerichtet. Jeder dieser Services konfiguriert eine Sammlung von Datenquellen als Metasuche. Im Fall von GEO-LEO gibt es einen Service mit den auf Seite XX genannten Quellen für die Metasuche auf der Startseite und einen weiteren nur mit den Beständen aus Freiberg und Göttingen für die thematische Suche. Die Konfiguration des verwendeten Protokolls, der Suchfelder und der Normalisierung wird für jede Datenquelle einzeln vorgenommen und kann in eigenen Dateien abgelegt werden. So lassen sich die Konfigurationen für häufig genutzte Server leicht wiederverwenden.

#### Weiterentwicklung

Pazpar2 wird von Index Data aktiv entwickelt. Gemeldete Fehler wurden häufig schnell behoben und über die „yazlist“ Mailingliste von Index Data können viele Fragen zur Konfiguration der Software mit den Entwicklern und anderen Nutzern diskutiert und geklärt werden.

## OAI Harvesting

Die in der GEO-LEO Metasuche eingebundenen Datenbestände „Digitalisate“, „Repositories“ und „Pangaea Data Publisher“ enthalten über OAI-PMH geharvestete Metadaten aus über 100 Dokumentenservern aus aller Welt.

Der verwendete Harvester wurde von Timo Schleier für [EROMM](European%20Register%20of%20of%20Microform%20and%20Digital%20Masters:%20http://www.eromm.org) entwickelt und zeichnet sich durch eine Weboberfläche zur Konfiguration aus. Mit der graphischen Oberfläche können die Fachreferenten direkt die inhaltliche Auswahl der zu harvestenden Daten treffen und sofort konfigurieren.

Nach Eingabe der URL einer OAI-PMH Schnittstelle lädt der Harvester die vom Server angebotenen Sets und zeigt sie zur Auswahl an. Sets, die seit dem letzten Bearbeiten hinzugekommen oder verschwunden sind werden dabei extra hervorgehoben. Eine Liste aktueller Fehlermeldungen in der Weboberfläche vereinfacht es, Probleme wie verschwundene Server (häufig geänderte Adressen der OAI Schnittstellen) zu erkennen und zu beheben.

(Bild Harvester.png / Filtern und Auswahl von Sets in der Weboberfläche des Harvesters)

Im Hintergrund laufen regelmäßig Skripte, die den Datenbestand von den konfigurierten OAI-Servern aktualisieren und die Daten in Solr-Indexe einbinden. Die Solr Indexe werden für die Metasuche direkt von Pazpar2 abgefragt.

## TYPO3

Der TYPO3 Server der SUB Göttingen stellt eine fertige Umgebung für beliebige TYPO3 Installationen zur Verfügung. Hier laufen unter anderem die drei an der SUB Göttingen betriebenen virtuellen Fachbibliotheken GEO-LEO, Library of Anglo-American Culture [\*](Lib%20AAC:%20http://aac.sub.uni-goettingen.de) und vifamath [\*](http://vifamath.de). Zur Präsentation der speziell in den virtuellen Fachbibliotheken vorhanden Inhalte gibt es TYPO3 Extensions:

* pazpar2: Einbindung der Metasuche
* nkwgok: Einbindung fachlicher Themenbäume
* ezbrequest: Darstellung von EZB Daten

(Bild: Verbindungen.pdf / Strukur und Abhängigkeiten der verschiedenen an GEO-LEO beteiligten Dienste)

### Pazpar2

Die *pazpar2* TYPO3 Extension stellt die Nutzeroberfläche für die Metasuche zur Verfügung. Sie kümmert sich um die Anzeige des Suchformulars auf der Seite und enthält die JavaScript Client-Software, die sich um die Anzeige der Ergebnisse und die Interaktion mit dem Nutzer kümmert.

Die Funktionen der Client-Software sind umfangreich. Sie beinhalten das Anzeigen der Ergebnisliste mit Blättern und Facettierung und die möglichst detailreiche Darstellung der Trefferdaten. Die Detailanzeige beginnt mit den elementaren Metadaten zu Titel, Personen, Serien und Zeitschriften und reicht bis zu speziellen Features wie der Verfügbarkeitsanzeige von Zeitschriften für die IP Adresse des Nutzers über den Journals Online & Print Dienst der ZDB, der Einbindung von Google Books falls dort vorhanden, der Einbindung Google Maps für Karten mit Geoinformationen in den Metadaten, dem Export der Metadaten im RIS oder BibTeX Format für Bibliographieprogramme oder einer Suche nach dem Titel im KVK.

Diese TYPO3 Extension kommt mittlerweile auch (angpaßt) in der Metasuche von CrossAsia [\*](http://crossasia.org/) zum Einsatz. Die Umsetzung einer Pazpar2 Suche für vifanord [\*](http://www.vifanord.de/) mit Einbindung in TYPO3 ist ebenfalls in Arbeit.

### Themenbäume

Zur Darstellung von Themenbäumen existiert die TYPO3 Extension *nkwgok*. Sie lädt Informationen zur Struktur der Themenbäume aus einfachen CSV Dateien – und alternativ aus lokalen Normsätzen im Pica OPAC.

Die CSV Dateien enthalten für jedes Themengebiet eine Zeile mit jeweils der ID des Datensatzes, dem deutschen und englischen Namen des Themas, der ID des Elternelements und einer zugehörigen Suchabfrage. Für die GEO-LEO Themensuche haben unsere Fachreferenten diese CSV Dateien erstellt und mit - teilweise komplexen – gleichzeitigen Suchabfragen in der Göttinger Online Klassifikation und Freiberger Dezimalklassifikation ausgestattet. Für die Suche im Geo-Guide existiert ein weniger granularer Baum.

Bei der Einbindung in die Site löst eine Themenauswahl sofort die Pazpar2 Suche über den jeweiligen Datenbestand aus. Die beiden TYPO3 Extensions sind an dieser Stelle gekoppelt.

Eine weitere Kopplung der beiden Extensions ist bei der Nutzung zur Anzeige von Neuerwerbungen möglich – die in GEO-LEO momentan nicht genutzt wird aber in Lib AAC, vifamath und den SUB Göttingen Seiten verwendet wird. Hier kann ein grober Themenbaum mit Abfragen definiert werden und es wird für jedes Thema eine Checkbox angezeigt. Nutzer können sich dann die Neuerwerbungen zu diesem Thema anzeigen lassen oder sie als RSS Feed abonnieren.

### EZB Abfrage

Zur Anzeige der „e-Zeitschriften“ ist die Extension *ezbrequest* im Einsatz, die die XML Schnittstelle der EZB nutzt und für vifamath entwickelt wurde. TYPO3 Extensions mit fast identischer Funktion gibt es auch an der SUB Hamburg (libconnect), an der HU Berlin (huubzeitschriftendienst: evifa) und der BSB München (bsbezb: vifaost).

### Piwik

Neben den selbstentwickelten Extensions kommt weiterhin die bereits existierende Extension *piwik* zum Einsatz, die eine Einbindung aller Seiten im TYPO3 System mit der Nutzerzählungssoftware Piwik [\*](http://de.piwik.org/). Durch den Einsatz derselben modernen Zählsoftware auf allen Webservern der SUB Göttingen erhalten wir eine realistische und konsistente Nutzerzählung.

## Open Source Als öffentlich finanziertes Projekt und großer Nutzer von Open Source Komponenten sind auch alle zur Umsetzung von GEO-LEO erstellten Programme und Konfigurationen frei verfüg- und weiterverwendbar. Sie sind mit allen anderen Projekten der SUB Göttingen im Account *subugoe* bei GitHub [\*](https://github.com/subugoe) abgelegt. Insbesondere:

* pazpar2 Konfiguration: https://github.com/subugoe/pazpar2-SUB
* OAI-PMH Harvester: https://github.com/subugoe/SUB-Harvester
* TYPO3 Konfiguration für GEO-LEO: https://github.com/subugoe/vlibs-typo3/tree/geoleo