Neue technische Entwicklungen auf dem Slavistik-Portal

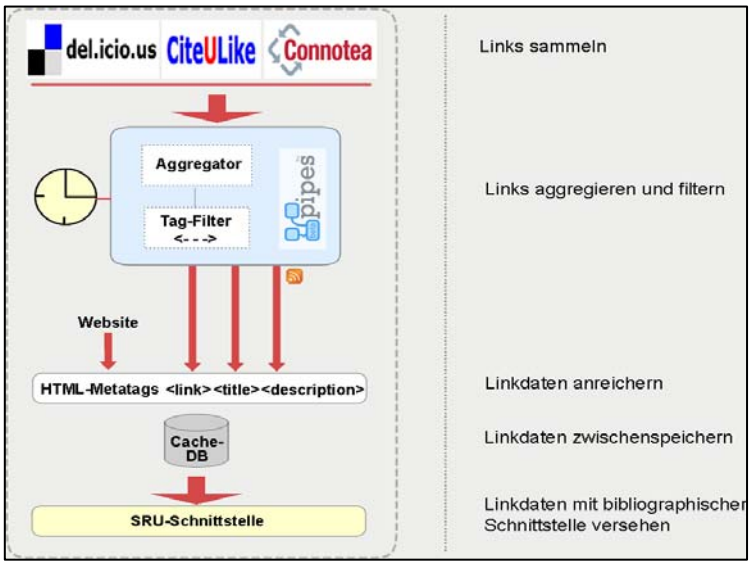
von Vladimir Neumann und Ivo Ulrich

Einleitung

Die Virtuelle Fachbibliothek Slavistik ist ein DFG-gefördertes Projekt, das an der Staatsbibliothek zu Berlin angesiedelt ist und dessen Ziel es ist, ein kostenfreies Portal für Slavisten im Internet – das Slavistik-Portal – aufzubauen. Das Projekt ist modular strukturiert und beinhaltet nach zwei Jahren Auf- und einem Jahr Ausbauphase nunmehr folgende Module: „Informationsführer Slavistik“ – sammelt relevante Internetquellen des Faches, „Neuerwerbungsdienst Slavistik“ – gibt einen Überblick über die Neuerwerbungen zum Sondersammelgebiet Slawistik der Staatsbibliothek zu Berlin, „Bibliographischer Datenpool“ – bietet retrodigitalisiertes bibliographisches Material zur deutschen und internationalen Slavistik, „Zeitschriften“ – gibt einen Überblick über die elektronischen und gedruckten Zeitschriften des Faches sowie ihre Zugangsmöglichkeiten, „Online-Tutorium Lotse-Slavistik“ – hilft den Slavisten beim Erlernen wissenschaftlicher Arbeitstechniken; und schließlich „Metasuche“ – das Herzstück des Portals, das eine simultane Suche in verschiedenen fachrelevanten Katalogen, Datenbanken und Bibliographien erlaubt.
Neben den Standardmodulen hat das Team der Virtuellen Fachbibliothek eine Reihe interessanter Entwicklungsarbeiten geleistet. Diese Arbeiten sollen hier kurz vorgestellt werden.

Weblink-Aggregator

Mit dem Weblink-Aggregator werden Internetquellen mit Hilfe der neuesten Web 2.0-Technologien zusammengetragen. Ausgangspunkt für diese Entwicklung war die Überlegung, dass eine ViFa nicht immer genug personelle Kapazitäten für das Auffinden von fachrelevanten Internetquellen besitzt. Gleichzeitig war die Tatsache bekannt, dass es im Internet eine große Fachgemeinschaft gibt, die auf den entsprechenden Webseiten Internetquellen sammelt. Mehr noch, die Internetquellen werden im Sinne des Social Bookmarking nicht nur gesammelt, sondern auch im Netz allen Interessenten zur Verfügung gestellt. Die bekannten Social-Bookmarking-Communities findet man z.B. bei Del.icio.us, CiteULike oder Connotea.



Vollautomatischer Weblink-Aggregator

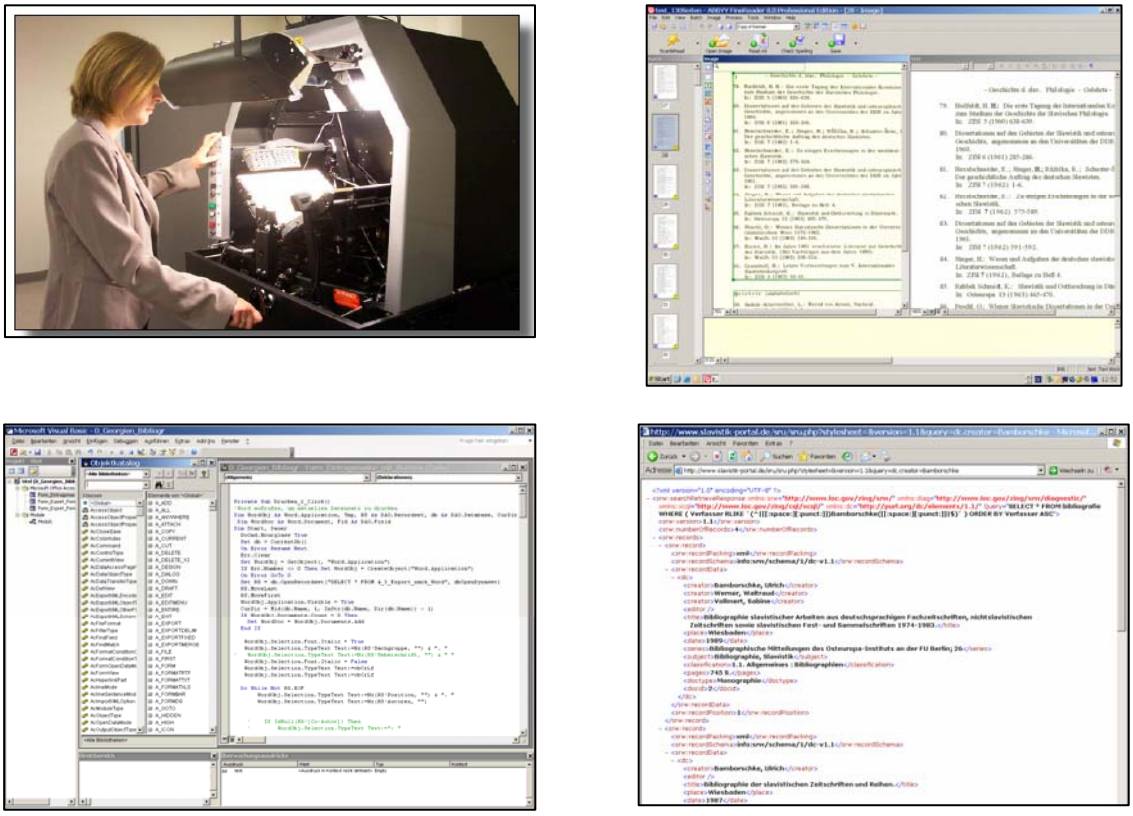
So entstand die Idee, die durch die Fachgemeinschaft beschlagworteten Links (sog. „Tagging“) zu sammeln, zu erschließen und im Portal zur Verfügung zu stellen.
Technisch sieht der ganze Vorgang wie folgt aus: Die Social-Bookmarking-Dienste werden über die Standardschnittstelle (in der Regel RSS) und nach bestimmten Tags wie „russische literatur“, „russische sprache“, „russian literature“, „русская литература“ u.ä. angesprochen. Deren XML-Antwort wird dann in einem Mashup-Dienst (Pipes™) nach bestimmten Kriterien gefiltert und in einer lokalen Cache-Datenbank abgelegt. Die Datensätze in der Cache-Datenbank werden mit den Metainformationen (Description, Keywords, Datum) in regelmäßigen Zeitabständen angereichert. Gleichzeitig erhält die Cache-Datenbank eine SRU-Schnittstelle (Searching und Retrieval via URL ist ein Nachfolgestandard von Z39.50), so dass sie als Target (maschinell ansteuerbare Datenquelle) in die Metasuche des Slavistik-Portals eingebunden und simultan mit anderen Datenbanken, Katalogen und Bibliographien durchsucht werden kann. Das Ergebnis dieses Verfahrens ist Folgendes: wenn ein Fachnutzer bei Del.icio.us eine Internetquelle zur Slavistik „bookmarkt“, wird diese Quelle mit zusätzlichen Metadaten ca. 2 Stunden später im Slavistik-Portal verfügbar und recherchierbar sein. In den letzten sechs Monaten sammelte die ViFa Slavistik auf diese Weise fast 2000 hochrelevante Internetquellen des Faches; täglich kommen drei bis sieben neue Internetquellen hinzu. Das Modul stellt ein effizientes Werkzeug für die Verstetigung der ViFa dar, weil es vollständig automatisiert (nur durch Technik gesteuert) funktioniert und dabei die Mitarbeit der weltweiten Fachgemeinschaft einschließt.

Halbautomatisches Konversionsverfahren für slavistische Bibliographien

Wie bekannt gibt es bis heute keine einheitliche und vollständige Bibliographie für deutschsprachige slavistische Veröffentlichungen. Um diesem Problem Abhilfe zu verschaffen, wurde das Modul „Datenpool Slavistik“ im Rahmen des Projekts entwickelt. Dabei handelt es sich um einen Sammelpool von bibliographischen Daten, die aus den deutschen gedruckten slavistischen Bibliographien in elektronische Form überführt sind. Die erste bibliographische Datenbank wurde aus der „Bibliographie slawistischer Veröffentlichungen aus Deutschland, Österreich und der Schweiz 1983/1987-1992“ (Hrsg. Von W. Gladrow, K. Gutschmidt, K. D. Seemann) konvertiert und in die elektronische Form überführt. Diese Datenbank ist nun seit zwei Jahren als „BibDatSlav“ im Internet erreichbar und durchsuchbar. Weitere Bibliographien werden folgen. Mit Unterstützung der Staatsbibliothek wurden vier Bibliographie-Reihen mit Hilfe eines Scan-Roboters gescannt, die einen Zeitraum von 130 Jahren der deutschsprachigen Slavistik abdecken. Bei der Auswahl der Bibliographien wurden sowohl die Publikationsart (Monographie, Zeitschriftenaufsatz) als auch der Ort der Publikation mit berücksichtigt (Westdeutschland, Ostdeutschland). Folgende Reihen wurden für die Konversion ausgewählt:

- Klaus-Dieter Seemann, Frank Siegmann: Bibliographie der slavistischen Arbeiten aus den deutschsprachigen Fachzeitschriften 1876-1963. In Kommission bei Otto Harrassowitz, Wiesbaden. Berlin 1965, 422 S. / Bd. II: (1964-1973), 736 S. / Bd. III (1974-1983), 745 S.
- Materialien zu einer slavistischen Bibliographie. Arbeiten der in Österreich, der Schweiz und der Bundesrepublik Deutschland tätigen Slavisten (1945-1963). Zusammengestellt von Irmgard Mahnken und Karl-Heinz Pollock. München, Sagner, 1963, 257 S. / Bd. II: (1963-1973) / Bd. III (1973-1983)
- Bibliographie slawistischer Publikationen aus der Deutschen Demokratischen Republik 1946-1967. Dem VIII. Internationalen Slawistenkongress gewidmet. Bearb.: Heinz Pohrt. Hrsg. vom Institut für Slawistik der Deutschen Akademie der Wissenschaften zu Berlin. Berlin, Akademie-Verlag, 1968, XVI, 400 S. / Bd. II: (1968-1972) / Bd. III: (1973-1977) / Bd. IV: (1978-1981) / Bd. V (1982-1986)

Die Voraussetzung der Übertragung einer gedruckten Publikation in eine Datenbank ist das OCR-Verfahren (OCR – Optical Character Recognition), bei dem Scans mit einer speziellen Software (hier ABBYY Finereader 9.0) in die Textform übertragen werden. Nachdem die auf diese Weise gewonnenen Daten auf mögliche Fehler überprüft wurden, müssen sie in eine strukturierte Form gebracht werden. Dies geschieht mit Hilfe einer Skriptsprache (in diesem Fall die Skriptsprache VBA – Visual Basic for Applications). Die Skriptsprache erlaubt es, die Daten nach vorprogrammierten Mustern aus dem durch das OCR-Verfahren gewonnenen Text zu extrahieren und zu strukturieren. Dabei wird darauf geachtet, die zu strukturierenden Daten soweit wie möglich zu zerlegen. Die größtmögliche Zerlegung erlaubt später eine Überführung des Datenmaterials in beliebige Datenbankformate. Im Fall der slavistischen Bibliographien werden die Daten in einer MySQL-Datenbank abgelegt, die über mehrere Schnittstellen (u.a. SRU) verfügt.



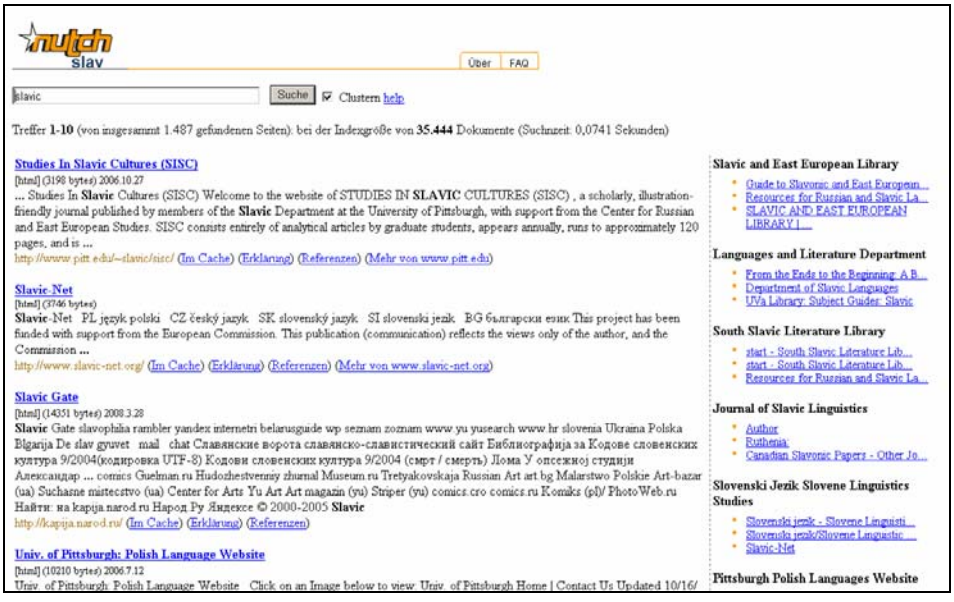
Prozess des halbautomatischen Konversionsverfahrens von slavistischen Bibliographien

Momentan ist ein Drittel des gescannten Materials (ca. 25.000 Datensätze) in die Textform überführt. Gleichzeitig wurde mit dessen Strukturierung begonnen.

Suchmaschinentechnologie „Nutch/Lucene“

Der Einsatz einer Suchmaschinentechnologie im Rahmen der ViFa Slavistik dient in erster Linie der Erweiterung des Contents mit dem Bezug zu slavischen Sprachen, Literaturen sowie slavischer Volkskunde.
Eine Suchmaschine (wie Google, Yahoo, Alltheweb u.ä.) sammelt in der Regel die Informationen aus dem World Wide Web, bereitet diese in Form eines Index intern auf und stellt diese Informationen über eine Suchmaske ihren Benutzern zur Verfügung. Die Technologie einer jeden Suchmaschine basiert auf drei Komponenten: Sammeln der

Webinhalte, Erstellen eines Index sowie Durchsuchen des Index über eine Suchmaske. Der Index, der den Kern einer Suchmaschine ausmacht, ermöglicht, dass viele Millionen (oder sogar Milliarden) Seiten mit Webinhalten (Webseiten, PDFs, Word-Dokumente, MP3, PowerPoint-Präsentationen, Zip-Archive usw.) in kürzester Zeit (meistens unter einer Sekunde) durchsucht und präsentiert werden können. Die Suchmaschine „Nutch/Lucene“ ist eine Suchmaschine im klassischen Sinn: sie sammelt die Webinhalte (über einen Crawler), bereitet sie in einem Index auf (sog. Lucene-Index) und ermöglicht dem Nutzer schließlich, die indizierten Inhalte komfortabel und schnell zu durchsuchen. Dazu wird „Nutch/Lucene“ unter General Public Licence betrieben und der Code steht als Open Source frei zur Verfügung. Die Suchmaschine wird durch eine große Gemeinschaft von Programmierern rund um den Globus kooperativ gepflegt und weiterentwickelt.
Im Zentrum des Interesses für den Einsatz von „Nutch/Lucene“ beim Slavistik-Portal stand das sogenannte vertikale Suchen (Vertical Search). Die Suchmaschinentechnologie bringt Werkzeuge mit, die es dem Team der ViFa ermöglicht, nur bestimmte Inhalte aus dem WWW herauszufiltern und in den Index aufzunehmen. Es handelt sich dabei um solche Inhalte, die einen direkten Bezug zur Slavistik haben.
Mittlerweile hat die ViFa Slavistik einen Prototypen dieser Suchmaschine aufgebaut, der auf den Servern des Projektes bzw. der Staatsbibliothek zu Berlin läuft. Im ersten Schritt wurden mehrere kleinere Indizes mit Inhalten aus ca. 40.000 Webseiten erstellt – alle mit dem Bezug zur Slavistik. Darüber hinaus ist der ViFa eine weitere Entwicklung gelungen. Da die Suchmaschine eine Reihe von standardisierten Schnittstellen mitbringt (z.B. OpenSearch) konnten diese Schnittstellen in die Metasuchmaschine (simultane Suche) des Portals eingebunden werden. Auf diese Weise kann der „Nutch/Lucene“-Index zur Slavistik auch über die Metasuche parallel zu vielen fachrelevanten Katalogen und Datenbanken durchsucht werden. Dadurch wird der Geschwindigkeitsgewinn der Suchmaschinentechnologie (Suchzeit unter einer Sekunde) auf die Metasuche übertragen, so dass die Nutzer an die gesuchten Fachinformationen sehr schnell herangeführt werden.

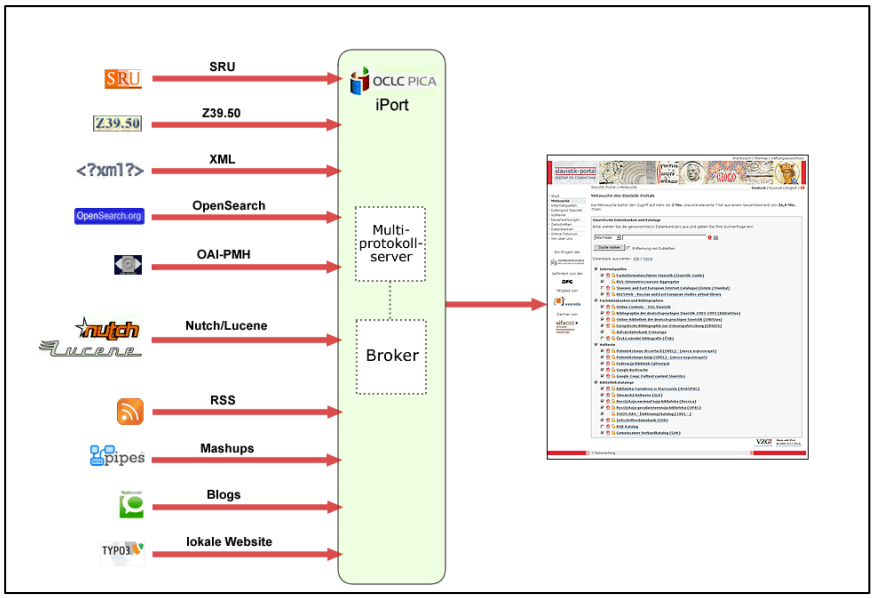


Prototyp „Nutch/Lucene“ des Slavistik-Portals

Im zweiten Schritt plant die ViFa Slavistik die Erstellung eines großen Index mit Bezug zur Slavistik. Der Prototyp-Index mit 40.000 Dokumenten wird einem Index mit 5 Millionen Dokumenten weichen (wobei die Suchzeit hier ebenfalls unter 1 Sekunde bleiben wird). In diesen großen Index sollen die Webinhalte aus allen Ländern Osteuropa einfließen, wobei das Relevanzkriterium stets der Bezug zur Slavistik ist.

Moderne Schnittstellen in der Metasuchmaschine des Portals

Die Metasuche des Portals basiert auf der Software iPort, die von der Firma OCLC/Pica entwickelt und gepflegt wird. In den drei Jahren der Laufzeit des Projekts hat das Team der ViFa auf der Suche nach Wegen zur Integration von slavistischen Katalogen, Datenbanken und anderen Quellen einige interessante Schnittstellen zum Einsatz gebracht. Die Schnittstellen erweitern die Funktionsfähigkeit der parallelen Suche und dienen der Erweiterung des Suchraumes. Neben den bibliothekarischen Schnittstellen wie Z39.50 und SRU kommen die Schnittstellen aus dem Dokumentations- bzw. Archivbereich wie OAI-PMH (Protocol for Metadata Harvesting) oder aus der Programmiertechnik wie XML (Extensible Markup Language) zum Einsatz. Des Weiteren nutzt die ViFa verstärkt Werkzeuge des Web 2.0 und wendet RSS (Really Simple Syndication), OpenSearch, Mashups, Blogs oder SRU-Lucene für die parallele Suche in bestimmten Datenquellen an.

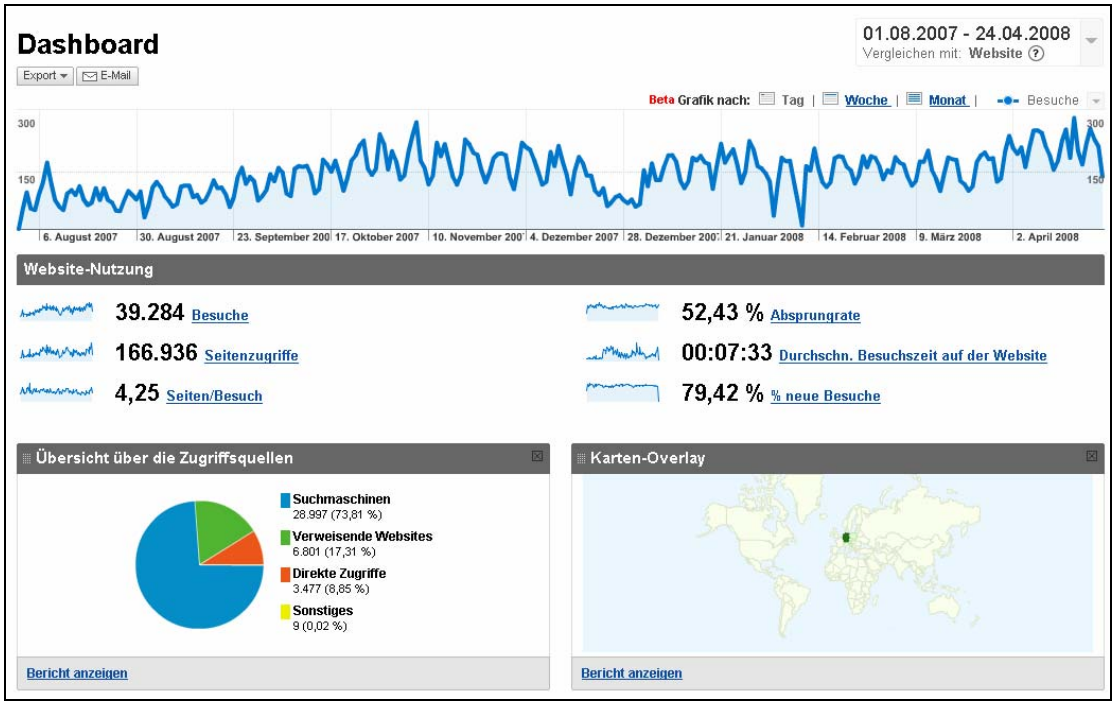


Moderne Schnittstellen in der Metasuche

Statistische Tools

Das Team der ViFa Slavistik legt seit Beginn des Projekts großen Wert auf den Einsatz statistischer Werkzeuge im Slavistik-Portal. Weil die ViFa ein Internet-Projekt ist, ist es wichtig, den „virtuellen Kunden“ zu kennen, der das Portal besucht und nach Informationen recherchiert. Auch vom Standpunkt der Usability aus ist es entscheidend zu wissen, ob der Nutzer seine Bedürfnisse auf der Seite befriedigt sieht oder ob er enttäuscht die Seite verlässt. Alle diese Informationen, die durch statistische Tools erfasst werden, dienen dem Team zur Verbesserung des Angebots und zu Korrekturen, um den Recherchewunsch des Nutzers schneller und effizienter aufzufangen.
Zur Messung der Besucherströme setzt das Portal die Software AWStats ein, die Server-Logfiles auswertet und zuverlässige Zahlen über die Portalnutzung liefert. Ergänzt wird das Statistikmodul durch einen selbstentwickelten sog. Tracker, der die Herkunft der Besucherströme sowie Art der Benutzeranfragen an die Module des Portals in Echtzeit visualisiert.
Das statistische Bild des Portals sieht wie folgt aus: in den ersten 3 Monaten nach dem Onlinegang von Mai bis Juli 2007 lag die Besucherzahl bei 6.000 Einzelnutzern (die IP-

Adresse der Nutzer wird nur einmal pro Stunde gezählt; die Mitarbeiter der SBB werden in der Statistik nicht berücksichtigt). Von August 2007 bis April 2008 haben das Portal 30.000 Nutzer besucht. Ab März 2008 liegt die Anzahl der Besucher bei ca. 200 pro Tag (was für eine kleine Fachgemeinschaft wie die der Slavisten sehr beachtlich ist). Zum Moment der Veröffentlichung dieser Publikation erreicht die Anzahl der täglichen Besucher ungefähr 400-450, wobei die Tendenz steigend ist. Es sind ca. 30.000 Metasuchanfragen im Jahr nach dem Onlinegang (vom April 2007 bis April 2008) verzeichnet worden. Die Durchschnittsverweildauer eines Besuchers auf den Portalseiten liegt bei ca. sieben Minuten, was auf eine intensive Nutzung der einzelnen Module schließen läßt. Die meisten Besucher gehen zum Fachinformationsführer Slavistik-Guide, was ein starkes Interesse an slavistischen Internetquellen impliziert, sowie zum Bibliographischen Datenpool, der retrokovertierte slavistische Bibliographien beinhaltet. Die meisten Besucher kommen aus Deutschland, den USA und Osteuropa.



Statistisches Bild des Slavistik-Portals vom 1.08.2007 bis 24.04.2008

Zusammenfassung

Es lässt sich sagen, dass die ViFa Slavistik mit der Entwicklung und dem Einsatz von neuen Modulen und Technologien, die über Standardmodule und -Technologien hinausgehen, bemerkenswerte Erfolge erzielt hat. Über den Weblink-Aggregator, der den Prozess des Auffindens und Katalogisierens von Internetquellen automatisiert, ferner das halbautomatische Konversionsverfahren, das eine große Menge gedruckten fachrelevanten Materials in eine elektronische, strukturierte Form überführt, und schließlich den Einsatz der Suchmaschinentechnologie „Nutch/Lucene“, der den Suchraum der Slavistik um mehrere Millionen fachrelevanter Dokumente bedeutend erweitert, hat das Team der ViFa Slavistik neue Technologien entwickelt und eingesetzt, sowie interessante Wege beschritten, die mit relativ wenig Aufwand von jeder anderen ViFa oder interessierten Fachgemeinschaft nachgenutzt werden können. Für den Erfolg dieser neuen Technologien sprechen insbesondere die Nutzungsstatistiken: über 400 das Portal intensiv nutzende Besucher pro Tag sind bei einer relativ kleinen Fachgemeinschaft, wie die der Slavisten, ein beachtliches Ergebnis.