GBV FAG Technische Infrastruktur

Protokoll der 38. Sitzung am 03.11.2009 in Göttingen

Teilnehmer: Aschenbach, Beilken, Blenkle, Borst, Busse, Hahn (Moderation), Mühl (Protokoll), Tuschick, von der Ahe

entschuldigt fehlen: Christof, Nord

Gäste (temporär): G.Steilen (VZG), S.Farrenkopf (SUB Göttingen)

Tagungsort und -zeit: SUB Göttingen, Gr. Seminarraum, 10:30 – 15.30 Uhr

I N H A L T

1	**Formalia**	**1**
2	**Verschiedenes**	**1**
2.1	Stand Teilnahme an der DFN-AAI	1
2.2	Anfrage der FAG LG	2
2.3	E-Journals: ERM und Statistik	2
2.4	Vergabe von URNs	2
3	**Touchpoint, Primo, Suchkiste, was kommt nach OPC4?**	**3**
4	**Nächstes Arbeitstreffen am 26.01.2010 in Bremen**	**5**
5	**Anhang**	**6**
5.1	ERM und Nutzungsstatistiken an der SUB Göttingen	6

1 Formalia

- Bestätigung der vorab vorgeschlagenen Tagesordnung bei Erweiterung um wenige Punkte
- Bestätigung des Protokolls der letzten Sitzung (von H.Busse)
- Nächstes Arbeitstreffen der FAG TI: 26.01.2010 in Bremen (Details siehe Topic 4)

2 Verschiedenes

2.1 Stand Teilnahme an der DFN-AAI

Thematisiert werden Probleme, die sich aus unterschiedlichen Nutzergruppen von Bibliotheken und Universitätsrechenzentren ergeben, die "Shibbolethisierung" wichtiger Campus-Services wie z.B. StudIP, Dokumentenserver, sowie die Frage der Qualitätssicherung der Identity-Provider-Daten. Unterschiedliche Player wie Uni-Rechenzentrum, Bibliotheken oder Staatsbibliotheken (mit externen Nutzern) haben unterschiedliche Klientel-Gruppen, sodass ev. die Regelung komplexer Zugriffsrechtestrukturen erforderlich sein kann. Die Einhaltung der (recht klaren) Richtlinien der DFN-AAI ist für eine Teilnahme an der Föderation basal. Allgemein wird der für eine "Shibbolethisierung" von Services erforderliche organisatorische Aufwand erheblich größer

eingeschätzt als der an sich schon nicht triviale für die technische Realisierung. Grundlegende organisatorisch-rechtliche Fragen wie z.B. die Universitätszugehörigkeit müssen in einem oft schmerzhaften Prozess geklärt werden. Nicht selten können z.B. einer Person mehrere Rollen mit unterschiedlichen, ev. auch widersprüchlichen Rechten zugeschrieben werden.

Der IDM-Connector soll nach einer Untersuchungsphase bei OCLC im November 2009 wieder für das Projekt in Jena freigegeben werden.

2.2 Anfrage der FAG LG

Im Protokoll des Arbeitstreffens der FAG Lokale Geschäftsgänge vom 11.09.2009 wird die FAG TI unter dem 3. Topic "Lokales Bibliothekssystem" um Mitarbeit gebeten: "Die Möglichkeiten zur Prozeßüberwachung durch den Systemadministrator sind in LBS4 gegenüber LBS3 eingeschränkt. Zu diesem Punkt soll die FAG TI einbezogen werden." (siehe https://www.gbv.de/cls-download/fag-lokale-geschaeftsgaenge/protokolle/2009-2011/FAGLG-5-30_110909.pdf).

Spezieller Hintergrund ist hier: Da im Gegensatz zur Realisierung im LBS3 im LBS4 ein Jaguar-Applikationsserver alle WinIBW-Clients über einen einzigen Prozess bedient, führt ein ev. für einen Client erforderlicher Neustart zum Service-Abbruch bzw. Neustart für alle weiteren laufenden Clients. Update-Prozesse sind allerdings nach wie vor eigenständig. Eine Änderung der Prozess-Architektur an dieser Stelle ist selbst bei kritischer Sicht sehr unwahrscheinlich. Herr Blenkle wird Hern Ahlborn fragen, wie die FAG Lokale Geschäftsgänge damit umgeht. Anschließend wird geklärt werden, in welcher Weise die FAG TI Stellung beziehen bzw. Empfehlungen an die VZG weitergeben wird.

2.3 E-Journals: ERM und Statistik

Die SUB Göttingen fragt nach aktuellen Strategien und eingesetzten Werkzeugen der Partner-Bibliotheken zur Bewältigung des wachsenden Arbeitsaufkommens im Bereich des Electronic Ressource Management (ERM) und der Aufbereitung von Nutzungsstatistiken. Einführend skizziert Herr Mühl die in diesem Bereich zunehmend schwierige Lage an der SUB Göttingen (Details siehe Anhang 5.1 "ERM und Nutzungsstatistiken an der SUB Göttingen").

Einzelne Rückmeldungen weisen darauf hin, dass Sachlage und Bewältigungsstrategie im ERM-Bereich an einigen Bibliotheken ähnlich wie in Göttingen sind und dass hier z.Z. ein erheblicher Mangel an Werkzeug-Unterstützung besteht. Einige Teilnehmer signalisieren Interesse an einer deutlichen Verbesserung. Das Arbeitsfeld wird allerdings als komplex und sehr in Bewegung eingeschätzt. Die Erwerbung von Lizenzen kann lokal hochpolitisch und von Vorlieben beteiligter Player (z.B. einzelner Professoren) geprägt sein. Nutzungsstatistiken werden an einzelnen Bibliotheken bereits in Form von COUNTER-Statistiken großer Anbieter zusammengetragen. Die Daten-Aufbereitung geschieht allerdings i.d.R. manuell via Excel und vergleichbaren Werkzeugen und ohne nennenswerte Integration. Eine gut integrierte Datenhaltung ist u.a. auch von Interesse für einen Vergleich der Nutzung von E-Journals im Open Access mit lizenzierten Angeboten. Auch im Bereich Bereitstellung einer integrierten Nutzungsstatistik besteht verstärkter Bedarf an standardisierten Formaten und Verfahren und an Automatisierung.

2.4 Vergabe von URNs

Der Fachbeirat thematisierte die Frage nach einer gemeinsamen Infrastruktur zur Vergabe von URNs. Primär zuständig für die Regelung der Vergabe von URNs ist die DNB. Will eine Bibliothek den Service der DNB nutzen, so ist sie auf eine für den Produktionsprozess einsatzfähige Resolving-Infrastruktur angewiesen, die lokale Links → URNs zuordnet. So vergibt z.B. das Goobi-Workflowsystem des Göttinger Digitalisierungszentrums (GDZ) z.Z. PURLs, die mittelfristig auf URNs gemappt werden. Die Rolle der VZG könnte in diesem Kontext sein, für

gehostete Digitalisate solche PURLs zu vergeben. Auf Verbund-Ebene ist die Sicherung der Stabilität und Zitierfähigkeit der Referenzen von Katalogeinträgen viel besser lösbar als auf Ebene der einzelnen Bibliotheken. Für eine langfristige Sicherung der Zitierfähigkeit elektronischer Ressourcen ist die Entwicklung einer gemeinsamen Policy sowie gemeinsamer Konzepte für brauchbare Resolving-Mechanismen durch die Verbünde erforderlich. Dies gilt insbesondere auch für die vielen kleineren Nachweissysteme, für die die Holder-Bibliotheken kein eigenes Resolving-System entwickeln können. Sinnvoll wäre hier ev. die URN nicht einrichtungsspezifisch zu vergeben, sondern über eine Art Treuhand-Verwaltung durch die Verbünde. Hierfür sprechen auch Bibliotheks-übergreifende Sammlungen, die besondere Erfordernisse bei der formalen Konstruktion von URNs stellen.

3 Touchpoint, Primo, Suchkiste, was kommt nach OPC4?

Zum Thema "Touchpoint, Primo, Suchkiste, was kommt nach OPC4?" stellt Herr Gerald Steilen (VZG) die bisherigen Erfahrungen aus dem seit August 2008 laufenden DFG-Projekt "Aufbau eines zentralen Nachweises und einer Recherche- und Serviceplattform für national lizenzierte elektronische Ressourcen" vor (WebBlog hierzu: http://blog.nationallizenz.de/2008/09/15/das-projekt/). Der bereits existierende Prototyp (http://finden.nationallizenz.de/) wird von der VZG auch kurz "Suchkiste" genannt. Diese basiert auf etablierter und leistungsfähiger Open-Source-Software, namentlich dem Indexierungswerkzeug Solr (http://lucene.apache.org/solr/), das Lucene (http://lucene.apache.org/java/docs/index.html) nutzt, und vufind (http://www.vufind.org/), einem modernen Portal-Framework für Bibliotheksressourcen.

Solr ist in der Lage in einem großen Datenpool Daten zu unterschiedlichsten Ressourcen und Quellen-Materialien zu integrieren. Es kann im Prinzip beliebige textuelle Daten indexieren und hierüber integriert suchbar machen, also neben bibl. Metadaten auch solche aus Volltexten und diversen anderen Quellen (z.B. jpeg-Metadaten). Sehr performant ist Solr bei Nutzung flacher Indexstrukturen, die via http-Schnittstelle aufgebaut werden können. Es gibt Solr-Instanzen die mit 500 Mill. Records umgehen können. Die hohe Performance und Ausfallsicherheit können nicht zuletzt durch die Clusterfähigkeit von Solr gesichert werden. Die "Suchkiste" enthält z.Z. rund 25 Mill. Datensätze und das komplette Nature-Archiv im Volltext. Der Index Aufbau ist schnell und in etwa mit PSI vergleichbar.

Erste Erfahrung der VZG weisen darauf hin, dass eine allgemein gültige Ranking-Konfiguration nicht sinnvoll möglich ist. Eigene Daten müssen je nach gewünschtem Ranking für Solr passend gewichtet werden. Für eine gewünschte Drilldown-Funktionalität muß das Mapping zwischen Begriffen selbst entwickelt werden, so etwa bei heterogener Sacherschließung. Die Einführung von Normdaten in Solr ist nicht trivial. Werden diese direkt bei der Indexierung den Records zugeordnet, so kann der Index sehr komplex werden und evtl. die Such-Performance ausbremsen. Werden sie erst bei der Ergebnis-Präsentation mit einbezogen, so sind sie nicht suchbar. Hier hat Oberhachingen noch keinen Lösungsvorschlag angeboten. Auch die Übernahme der lokalen Sacherschließung oder anderer "lokaler Spezialitäten" gestalten sich evtl aufwendig. Vorarbeiten hierzu sind im Gange, eine Unterstützung von Pica bzw. aus Oberhachingen ist erforderlich.

Solr wird der zukünftige GBV-Standard in Sachen Indexierung mit TouchPoint sein. Gegenüber dem sicherlich sehr performanten PSI hat es den großen Vorteil der Transparenz des Quellcodes, der großen Anwender-Community und dass man mit Web-Entwickler-Knowhow sehr weit komme. Eine Stärke von Solr ist, dass es bereits in der Grundausstattung umfangreiche Funktionalität und Konfigurierbarkeit bietet (z.B. Facettierung über jedes beliebige deklarierte Feld). Um die umfangreichen Möglichkeiten nutzen zu können müssen sich Anwender bzw. Entwickler mit ihnen vertraut machen. Einschlägiges KnowHow ist also unabdingbar. Support gibt es hier aber durch

eine große Anwender-Community incl. Kommunikationsplattform und Mailinglisten (http://lucene.apache.org/solr/mailing_lists.html). Solr-Bugs werden i.d.R. in vergleichsweise kurzer Zeit von den Entwicklern beseitigt. Die VZG kann Hilfestellung beim Einsatz des Solr-Kernsystem anbieten, etwa beim Aufbau eines lokalen Solr-Index. Ingest-Prozesse sind über die http-Schnittstelle möglich, aber auch direkt über (eigene) Java-Programme. Search-Requests können über http ausgeführt werden. Solr selbst bietet über eine http-Schnittstelle die Funktionalität der Lucene Bibliothek an. Für Einsteiger wie auch für Fortgeschrittene empfiehlt Herr Steilen die Lektüre des erst im August 2009 ersten veröffentlichten Buches zum Thema: "Solr 1.4 Enterprise Search Server" von David Smiley und Eric Pugh (Verlag Packt Publishing; ISBN-13: 978-1847195883). Empfohlen wird auch das Online-Tutorium des Lucene-"Erfinders" Eric Hatcher (http://www.youtube.com/watch?v=eRQeYiuPgMA).

Solr bietet umfangreiche Funktionalität im Bereich Indexierung und Suche. Dies gilt aber nicht für die oft vor dem Import erforderliche Konversion eigener spez. formatierter Daten. Hier können Community-Mitglieder nach (Eigen-)Bedarf Werkzeuge entwickeln, die i.d.R. allen Anwendern zur Verfügung stehen. So existiert z.B. das Import-Tool SolrMarc, das Import und Indexierung von Marc-formatierten Records in SOLR ermöglicht (siehe http://www.oss4lib.org/node/560 bzw. http://code.google.com/p/solrmarc/). Verbund-Kataloge können solch ein Werkzeug unmittelbar nutzen, da sie marc21 unterstützen. Die VZG plant über die Bereitstellung einer spez. Ingest-Routine den direkten Import von Pica+-Daten in Solr zu ermöglichen. Für eigene Daten wird laut Herrn Steilen eigene Expertise in Sachen Konvertierung / Mapping wichtiger denn je. Aus der FAG TI gab es den ergänzenden Hinweis auf den Relaunch des WiWi Fachportals, das auf Solr umstelle und bei dem die Vorab-Datenkonvertierung die meiste Arbeit mache. Gegebenenfalls wird laut Herrn Steilen die Entwicklung oder Anpassung von Werkzeugen für spez. Datenbestände unumgänglich sein. Eigenes KnowHow der Programmiersprache Java ist hier sehr vorteilhaft, da Solr, Lucene sowie die Tools aus der Community in Java geschrieben sind.

Aktuelle Portaltechnologien wie das hier eingesetzte vuFind, aber auch Touchpoint (OCLC) oder Primo Central (exlibris) setzen direkt auf der Indexmaschine Solr auf. Das Recherche-Interface der "Suchkiste" basiert auf vuFind. Für die Suchkiste selbst wurde von Prof. Dr. Bekavac (http://www.fh-htwchur.ch/sii-home/team/dozenten/bekavac-bernard/) von der HTW Chur (Schweiz) ein Oberflächenkonzept (Tabellen zur Ergebnispräsentation und Navigation) entwickelt. Im Januar 2010 wird die produktive Version dieses Oberflächenkonzepts freigeschaltet. Die "Suchkiste" bietet zum Such-Einstieg nur eine Standard-Suche (einfacher "Suchschlitz"). Auf eine erweiterte Suche wurde verzichtet, da Untersuchungen zeigen, dass letztere nur minimal genutzt wird. Dennoch ist das Ziel, dass ein Suchinterface den Bedürfnissen Studierender ebenso gerecht wird wie denen von Bibliothekaren. Vorgabe ist ein freier Zugang zur Recherche über alle Daten. Primär wird eine Relevanz-orientierte Sortierstrategie unterstützt, die nicht nur den einen exakten Treffer sucht, sondern auch auf große Freiheitsgrade zur weiteren Auswahl in bereits erhaltenen Suchergebnissen setzt, sodass z.B. eine schrittweise Eingrenzung von Ergebnis-Sets durch Facettierung oder ähnliche Treffer möglich ist. Das Ranking in der Präsentation der Ergebnis-Sets ist flexibel konfigurierbar, sodass z.B. Treffer in Titeln oder Volltext Priorität erhalten. Unterschiedliche Zugriffrechte beeinflussen sowohl das Relevanz-Ranking als auch die Trefferpräsentation.

OCLC Deutschland wird eine Kombination von Touchpoint und Solr im OCLC-Portfolio als "Standard“ anbieten. Eine von Oberhachingen ausgelieferte Solr-Version wird allerdings nur das Kernsystem und keine spez. Werkzeuge einschließen. In Bochum ist eine Touchpoint-Installation auf der Basis von Solr im Einsatz. Herr Rathai ist offizieller Projektleiter für TouchPoint im GBV. Neben Solr und vuFind gilt auch für Touchpoint, dass Java-KnowHow für Anpassungsarbeiten hilfreich ist. Der Quellcode liegt zwar nicht offen, Touchpoint nutzt aber den Servlet-Container Tomcat, der die Einbindung eigener Java-Klassen erlaubt.

Primo Central hat laut Herrn Steilen neben den hohen Kosten den Nachteil, dass es ein reines Recherche-System ist. So wird z.B. die Ermittlung der Verfügbarkeit von Ressourcen nicht

unterstützt. Dies ist zwar z.Z. bei Touchpoint nicht anders, es gibt aber schriftliche Vereinbarung mit OCLC zur Nachbesserung. Ein Vorteil von Primo hingegen ist die recht ausgefeilte, integrierte Statistik. Wie Serial Solutions baut es einen sehr umfangreichen zentralen Solr-Index aus erworbenen Datenbeständen auf, in den auch eigene Daten integriert werden können. Auch für die VZG ist eine analoge Lösung mit Solr durch Aufbau eines großen zentralen Index denkbar. Eine föderierte Suche auf dezentralen Solr-/Lucene-Indizes ist bisher nicht vorgesehen. Diese ist prinzipiell machbar, allerdings nur unter der Voraussetzung einheitlicher dezentraler Index-Strukturen.

Das Geschäftsszenario für interessierte Bibliotheken ist, dass diese ihre Verträge immer mit Oberhachingen abschliessen, also nicht mit der VZG. Die Rolle der VZG sieht Herr Steilen in einer "Starthilfe" beim Einsatz der neuen Technologien, nicht aber in der Übernahme der Maintenance der Produktivsysteme. Aus der FAG TI kommt der Vorschlag, dass kleinere Bibliotheken, die auf Grund begrenzter Ressourcen nicht über Kapazitäten für Eigenentwicklungen verfügen, Support durch die Normierung von Datenstrukturen (Indizes, Import) erhalten sollten. Unabhängig von der Einführung der neuen Technologien bleibt laut Herrn Steilen der klassische OPAC auch zukünftig integraler Bestandteil des Pica-Bibliotheksystems.

Abschließend plädiert Herr Steilen für eine verbesserte Absprache zwischen Bibliotheken zur Klärung des gemeinsamen Bedarfs. Zu oft werden einzelne spezielle Fragen an die VZG herangetragen, ohne dass Hintergrund und Kontext klar sind. Nachdrücklicher Wunsch von Herrn Steilen sind (schriftlich fixierte) Anforderungsspezifikationen, die den gemeinsamen Bedarf der Bibliotheken nach Priorität und integriert zusammenstellen.

4 Nächstes Arbeitstreffen am 26.01.2010 in Bremen

Die FAG TI wird am 26.01.2010 in Bremen tagen. Als mögliche Topics für Bremen oder anschließende Arbeitstreffen werden vorgeschlagen:

- Vorstellung der Portaltechnologie Primo der Firma Exlibris
- Vorstellung des Digitalisierungspaketes der VZG / Goobi - hier bietet sich der Folgetermin in Göttigen an
- Übersicht Ebookportale / Erfahrungen in den Standorten - Herr Busse hatte das thematisiert
- Digitale Semesterapparate
  Was bietet der Verbund als Support / Infrastruktur ?
- VZG-Support für Hochschulbibliographie

  → siehe auch "Hochschul- und Regionalbibliografien" im GBV-Wiki
  (http://www.gbv.de/wikis/cls/Hochschul-_und_Regionalbibliografien)
- Zielvereinbarungen und Strategiepapier: Worum soll sich FAG TI kümmern ?
- Aktuelle und mögliche Lösungen der Bibliotheken zur LZA
- Erfahrungen mit MyBib edoc und ein Vergleich zu Subito
- Themen für die kommende Verbundkonferenz

5 Anhang

5.1 ERM und Nutzungsstatistiken an der SUB Göttingen

Einleitend zum Topics 2.3 "E-Journals: ERM und Statistik" (siehe oben) skizziert Herr Mühl die schwierige Lage an der SUB Göttingen wie folgt:

- An der SUB Göttingen (wie vermutlich an einigen GBV-Bibliotheken) wuchs im Laufe der letzten Jahre der Arbeitsaufwand zur Bewältigung der vielfältigen Aufgaben insbesondere im Bereich Lizenz-Erwerbung und -Verwaltung von E-Journals und Online-Datenbanken erheblich an. Teilweise rasant steigende Kosten von E-Ressourcen (insbesondere im naturwissenschaftlich-technischen Bereich) bei gleichzeitiger Reduktion der Finanzmittel erfordern die jährliche Revision des Lizenz-Bestandes. Für die Entscheidung über An- und Abbestellung müssen viele Uni-Facheinrichtungen in mehreren Evaluationsrunden intensiv einbezogen werden, da neben zentralen Finanzmitteln der Bibliotheken zunehmend Eigenleistungen der Fakultäten in die Finanzierung der Ressourcen einfließen. Aufgrund des Sparzwangs erwarten die dortigen Entscheider vermehrt zuverlässige Statistiken zur Ressourcen-Nutzung durch die eigene Einrichtung und die anderer Fachbereiche. Zunehmend wird die nutzungsabhängige Verteilung der Kosten lizenzierter Ressourcen gefordert.
- Die jährliche Revision eines Bestandes von vielen Tausend (!) Ressourcen schließt das Einholen der diversen Angebotsvarianten in Einzel- und Paket-Lizenzen und ihrer Preise sowie deren Zusammenstellung und Optimierung in Varianten von Gesamt-Bestellszenarien ein. Kosten-Vorteile aus Paket-Angeboten, gemeinsamen Prioritäten der Fachbereiche oder kostenfreier konsortialer Nutzung ("cross access") stehen hier den Wahl-Beschränkungen durch Konsortial-Verträge (z.B. Sperrung Print-Abbestellung) und anderen Verpflichtungen oder Interessenlagen gegenüber.
- Diese sehr komplexen Aufgaben müssen mit unzureichenden Mitteln bewältigt werden. Schon die realen Lizenz- und Kostenverhältnisse insb. von Paketstrukturen lassen sich im genuinen Pica-LBS-Werkzeug ACQ nicht angemessen abbilden und schon gar nicht auswerten (gilt anscheinend auch noch für ACQ4!). Zur Evaluation müssen Daten aus unverbunden nebeneinander stehenden Datenpools (Pica-CBS/-LBS, EZB, Lizenz- und Preislisten der Anbieter, eigene Listen, Prioritätslisten der Fakultäten) integriert werden. Die zentralen Werkzeuge ACQ und EZB erweisen sich als zu monolithisch oder unvollständig in ihren Möglichkeiten reale Sachverhalte und Abhängigkeiten abzubilden (EZB: Lizenzstruktur - ACQ: Kostenstruktur). Das Daten-Mapping zwischen EZB und ACQ ist nur mittels verschiedener Identifier auf Umwegen und allenfalls unvollständig möglich. Wichtige Daten wie z.B. Verlagsangaben sind in der ZDB bzw. im Pica CBS nicht normiert, sodass eine Daten-Selektion nach Verlag oft nicht zuverlässig machbar ist. Notgedrungen muss auf "altbewährte" einfache Werkzeuge wie MS-Excel und MS-Word zurückgegriffen werden. Die erforderliche Teamarbeit auf verteilten und wenig integrierten Daten führt unvermeidlich zu einem deutlichen Mehr an Organisationsaufwand, Informations- und Datenaustausch sowie Fehlern. Die Folge ist auch ein ausgeprägter Mangel an Transparenz im Bereich von Lizenz-Angeboten, -Bestand, –Kosten und Verbindlichkeiten für die beteiligten Player (insb. Sachbearbeiter, Fachreferenten, Kosten-Controlling, Benutzung).
- Eine angemessene Infrastruktur zur Auswertung der Nutzungsstatistiken fehlt komplett. Eine deutliche Verbesserung der Lage durch automatisiertes Abholen von Nutzungsstatistiken im COUNTER-Standardformat (http://www.projectcounter.org/) ist inzwischen machbar, zumindestens bei großen Anbietern, da diese z.T. bereits eine SUSHI-Schnittstelle (NISO-Standard Z39.93; http://www.niso.org/schemas/sushi/) hierfür anbieten. Allerdings fehlt auch hier noch ein zentrales integrierendes Werkzeug (Stichwort Business Intelligence (BI)).