Summary of the 3rd German VuFind Community Meeting in Frankfurt am Main, Germany

Day 1: 2014-Sep-24

Uwe Risch (head of the union center of HeBIS, the library union of the State of Hessen at Frankfurt am Main) welcomes all participants. Bettina Sunckel (HeBIS IT) takes over the introduction and reports that, sadly, Uwe Reh, who was very much involved with the preparation and organization of the meeting, is to be excused for being ill.

Bettina Sunckel then features the **first session of presentations**.

The first speaker is **Robert Strötgen** (Georg-Eckert-Institute for International Textbook Research (GEI), Braunschweig) with a presentation created with his colleague Jessica Drechsler about TextbookCat, the new search tool for the collection of textbooks at GEI. Thereby, Robert follows up Jessica's presentation at the previous German VuFind Community Meeting, which announced their project as targeted.
The reported challenges of textbook retrieval are their commonly meaningless title and parallel regional editions which makes the search especially difficult. In order to supplement the existing opac with a user-friendly intuitive faceted search for textbooks, an easy to implement open source solution was required, which could be discussed within a big community of users. VuFind was chosen, and the implementation was possible as part of a government-funded project.
The GBV central index had to be suited by an additional solr field in which all data from subject cataloging, prefixed by an indicator of the data category, are indexed, additionally prefixed by the local library number. To enable useful facetting by categorized local metadata, from this new coarse-valued facet field, refined values are derived by filtering on the local library number and by mapping the remaining field value by the prefix indicator onto several differentiated labels as virtual facettes. The VuFind configuration file facets.ini was edited and the virtual facets were defined as translateable.
Unfortunately, with this deep metadata analysis, erroneous indexed codes now become evident in the facets, such that continuous data maintenance is even more required.
Since the public beta test started in May 2014, an evaluation of the new facetted search has been carried out and will be repeated regularly in order to align TextbookCat to the user satisfaction. Further plans cover an upgrade to VuFind 2, usability enhancements, and, in the long run, the integration of international textbook catalogs.

The subsequent discussion focuses on the importance of introducing hierarchical facets into VuFind.

As second speaker, **Gebhard Engelmann** (UB Magdeburg) relects the processes which are necessary in order to establish a discovery system. They are based on the analysis of the experience during the implementation of UBfind, a joined project of UB Magdeburg and the GBV service center. Seven processes are presented: decision-making for having a discovery service,

choice of certain products and components, analysis of the determining factors, design, launch, maintenance, and participation of the staff.
Not only some librarians, who's notion of bibliographic search was not in accordance with the now imprecise result sets, the result sorting according to non-formal criteria, and the blurred classificatory facetting, have been dissatisfied with the paradigm shift from opac to discovery service but also some users, who had already built up a strong opac competence.
The staff acceptance has to grow with joined project work, the users are increasingly content.

Afterwards, as third speaker, **Stefan Winkler** (BSZ Konstanz) first summarizes relevance ranking as basic quality and then describes it in the context of merging result sets from different sources. The BSZ union catalog is used cooperatively by 1200 libraries for cataloging and a full dump of this catalog is indexed using Solr. Libraries can get a Vufind template with a specialized view of the whole Solr index, restricted on their holdings. Ten libraries already use it this way, either in beta or in production mode.
The first part of the presentation shows how difficult the handling and the traceability of any certain estimation after the configuration of many steps of data processing and evaluation are in practice to approach the subjective measure of the relevance of search results.
The second part portaits different approaches to process multiple result sets: Filtering within the same Solr index, boosting parts of an index, sharding multiple Solr indexes of similar or different Solr version or schema, SolrCloud, Solrmarc and Beanshells, as well as SolrFusion.
Concluding recommendations concern, above all, the disclosure of the ranking procedure, the avoidance of merging result sets if possible, for instance by seperate tabs or own indexes, as well as the alignment of structure and values of the index fields and relevance scores from different sources.
From the subsequent discussion, different opinions arise about considering statistics of usage for boosting result documents: on the one hand, the recommendation of important results may be also useful for others, on the other hand, the most used media should not constantly constitute the first results. On the one hand, textbooks are important search results in a subject area, but on the other hand, they might be redundant because they are recommended already in the lectures. Thus, one should be clear on what to consider as relevant results for the users and decide accordingly.

As forth presentation, a summary of the projet finc follows on the occasion of its approaching completion, reported by **Leander Seige** (UB Leipzig).
November 2014, the EFRE-funded project finc will be completed after 3 years. During the project, for 11 Saxon university libraries a ressource discovery service has been implemented according to individual requirements. Each library has received an individual VuFind front-end of the central Solr index, and additional enrichment data as well as the ExLibris PrimoCentral index have been included.
The combination and output of the joined result list from over 30 heterogeneous data sources is achieved by using PCBridge and by mapping index fields and aligning the boosting.

The integration of authority data has succeeded. The facets offer the possibility to lift restrictions and to broaden the search. The patron-driven acquisition can include the indexed booktrading data, the licensing of publications performs seamlessly in the background, and orderings for printed publications can be triggered.
On September, 18, 2014, the finc users' pool ("Nutzungsgemeinschaft") was founded as nonprofit association, and the joint institutions raise money for the maintenance of the infrastructure. One of the latest achievements is the use of SolrFusion. Also, un upgrade to VuFind 2 is pending.

Q: During the subsequent discussion, there is a question concerning the costs to be expected by the users' pool. A: As cost model, to a basic amount additional features as contracted are calculated as well as the metadata to be found in the catalog, such that for bigger institutions, the larger requirements of infrastructure usage like machine load and error treatment are applied.

After a break, Bettina Sunckel continues to lead the **session of the second half of presentations**, which focus on technical aspects.

Magdalena Roos (VZG Göttingen) in the fifth talk provides information about PAIA (Patrons Account Information API), a manufacturer-independent interface for user accounts and loan processes. PAIA defines an open interface, the PAIA server receives HTTPS requests in JSON format, translates them into requests to the library software and responds with JSON data. Authentication at the user account and password change are executed by the module PAIA auth. The module PAIA core allows for the account and charges view and all ordering interactions. PAIA 1.x communicates with LBS3. Since June 2014, their PAIA server is productive. Instructions for the setup of a PAIA server can be found at https://www.gbv.de/Verbundzentrale/serviceangebote/paia-service.
For VuFind, a PAIA driver has been developed, which is delivered as a standard with VuFind 2. The presentation concludes with an outlook on PAIA 2.0 with full LBS4 support.

Q: A question concernes the cost model. A: The setup of the PAIA service costs 630 EUR (one daily rate); the annual maintenance expense arise out of the staff size.

As sixth presentation, a report on work in progress follows by **Martin Czygan** (Uni Leipzig), composed together with Anke Hofmann, about the integration auf authority data and other data sources into a discovery system, concerned from raw data up to the services.
The issue is part of the project finc. Authority data and their link-up form a network of facts, of which little can be seen in the catalog. The facts are widespread and not necessarily accessible via a web interface. Such facts can turn out to be, e.g., biographical data of a person, like lifetime dates and spans, workplace oder profession. They can be obtained directly from GND or using APIs like http://lobid.org/api. Wikipedia references to GND can also be used from

DBPedia.
Possible queries like “Which persons of equal profession were born to the same city?” can be answered. Certainly, also entities besides persons can be specified.
Challenges arise from the amounts of data - at least several hundreds of millions of triples - as well as from the use of adequate query languages.

Q: A question from the audience refers to the application of such queries at the search user interface and the usability optimization. A: Martin Czygan reports that during the next weeks the operational test will produce some feedback.
Q: Additionally, the opinion is expressed, that the possibility to build a knowledge portal with VuFind surely exists, but libraries should concentrate on an intuitive search for publications and avoid drowning their discovery systems in additional features.

Filipe Bento (EBSCO) follows up as seventh speaker with his presentation "Embracing OpenSource: Introducing EBSCO Discovery Service’s VuFind Module" and points out how the open-source community is supported by EBSCO by adapted modules for VuFind, Koha, Moodle, and others. Furthermore, he reports how EBSCO is looking actively for experience from the community in order to consider the requirements for improvement and further development for EBSCO’s tools.
Concluding, an insight into the current VuFind 2.0 Module is shown, which has been published by EBSCO in cooperation with the VuFind community. A VuFind 1.3/1.4 demonstration can be found soon at http://vufinddemo.ebscohost.com; version 2.3’s demonstration site will soon be made available. The slides provide a detailed description of EDS integration for the several VuFind versions.

Q: Afterwards, the question is posed, whether the module for VuFind 1.3 will be maintained. A: The answer assures that EBSCO will go on developing first for version 1.3 and then for version 2.3. Version 1.x will not be dropped.

Andreas Koch (outermedia GmbH, Berlin), as eighth speaker concluding the day’s agenda, is talking about SolrFusion and sharding across heterogeneous index structures.
If several Solr instances are to be queried, and the search results be presented as one integrated result list, one faces a problem, if the indexes differ structurally. If an alignment of the Solr schemas and re-indexing is not possible or not appropriate, the open-source software SolrFusion can be the solution. SolrFusion results from a subproject of finc and has been developed together with the University of Leipzig.
SolrFusion is a Java web application with XML configuration and extensibilty via plugins. There are no further dependencies, and VuFind requires no costumization.
The Fusion Schema containes a translator for the communication between VuFind and the several Solr instances. Transformation rules define the interpretation of the structured VuFind request to requests to the respective Solr instances. The incoming Solr responses are integrated and send to VuFind.

The most important Solr features like eDisMax and DisMax syntax as well als highlighting, facets, MLT (more like this), and request parameters are supported. Duplicate result documents can be merged by a configured Solr field with prioritization. Faceting can be mapped to virtual facets. For integrated relevance ranking, a plugin can be implemented.

Questions concern, above all, the integrated ranking: The score values can be scaled by predefined factors. A merge across several fields (e.g., author, year, isbn) is generally possible and can be implemented by a plugin.

The first day closes with the anouncement that, unfortunately, the next day's talk about migration from VuFind 1.3 to 2.x by Markus Beh has to be canceled. Instead, a discussion about the VuFind upgrade is offered.

Day 2: 2014-Sep-25

The second day starts with Bettina Sunckel introducing the **first session of presentations**.

The first speaker is **Oliver Goldschmidt** (TUB Hamburg-Harburg) about Tabs and printed material, and how to integrate an external index into TUBfind.
TUBfind is a Discovery-System, which had been introduced four years ago and which was one of the first VuFind based systems in Germany. This year, TUBfind was extended by PrimoCentral as an external index. TUBfind is using the index from the Common Library Network (GBV), which includes bibliographic data form the Common Library Network, and the PrimoCentral index; since the results from PrimoCentral have to get transformed before they can get used in VuFind, TUBfind is using the PrimoBridge (which is delivering a result list similar to Solr). The TUBHH chose to display the results from PrimoCentral and from the GBV in two separate tabs (although the Bridge would be able to merge the results into one list), because there are several problems with the blended list: Facets are structurally and in their content completely different to each other, so that its not possible to find a Boost value which is valid for all results in PrimoCentral.
Since PrimoCentral only contains electronic publications its probably good for users to get a link to the printed equivalent with the electronic PrimoCentral match. This is how it works in a nutshell: The GBV index is queried again with the relevant metadata from any PrimoCentral match, to determine, if it is available printed in the library. Having a result, this is being analyzed to find a correct match and the ID of the printed record. This ID is displayed as a link to the record in TUBfinds GBV-tab. But there are still some problems, which are preventing a unique match: In some cases the publishing year of the ebook is different from the publication year of the printed book in the GBV index or the year is completely missing in PrimoCentral. Multiple journal volumes in one book can also be a problem, as well as when a journal volume is divided on multiple books.

One question in the discussion is concerning the finc project: What do they do about the problem of different scoring algorithms? This was an iterative process, in which the boosting value has been adjusted very carefully, until it was considered to be a good value for most of the use cases.
A suggestion about merging facets of both result lists is to map the values using accordance tables. This is difficult for classification facets, because most of the matches do not have a class.
A question about experiences with the Two-tab-solution cannot be answered, because serious statistics are missing until today.

As second speaker, **Jan Frederik Maas** (SUB Hamburg) is reporting about beluga and the usability optimization of a consortial discovery system.
In the project beluga, one of the main requirements consists in maximizing the user satisfaction with scientific research, particularly with regard to the participation of the users in the development process. The adoption of VuFind and the integration of the PrimoCentral index have been accompanied by two usability studies. Therefore, a heuristic evaluation concerning 94 criteria established for search-engine evaluation has been conducted. Two groups of students have been checking the criteria, the results were weighted and summarized. In a second step, user studies have been executed using the thinking-aloud method, in which users had to solve given tasks while using the system. Furthermore, screenshots have been employed to evaluate, which user-interface elements were important, insignificant, or incomprehensible.
From the studies, numerous modifications resulted, e.g., to prevent frustration at zero hits by helpful links and to utilize spelling suggestions from the requested part of the index. Other improvements concern, e.g., alphabetical sorting of the facet values for subject and author instead of sorting by frequency.
The second study consisted of the implementation and evaluation of a very user friendly interface, which presented the search results of two separate indexes. For the two result lists from the GBV- and the PrimoCentral index, the evaluation lead to the choice of the two-tab solution in contrast to the integrated result list and the parallel presentation, especially because of the easier relevance ranking and the heterogeneity of the index data (facets, types of work). The launch of the new version with Primo integration and the accompanied update to VuFind 2.3 are pending until the end of 2014.

Complementary to the usability studies, the question is raised, whether search loggings are evaluated. This is answered by the statement that for a standardized assessment by means of user studies with subjects, a clear setting is preferred because the interaction with the system can be tracked better.
Also, the proposed alphabetical sorting of certain facet values gains interest, and a question is targeting at further suggestions from the participating subjects for the improvement of facets. However, this was not the case, but the alphabetical sorting again does not seem optimal to a part of the audience in consideration of the many subject- and author values. It seems worthwile with few facet values, though.

After a break, Stefan Winkler announces the offer to let the **4th German VuFind Community Meeting take place in the south-west of Germany** - the location could be for instance Stuttgart, Freiburg, or Konstanz. The proposal is appreciated very much by the participants.

Subsequently, the open program follows, featured again by Bettina Sunckel. It starts with a discussion about the **upgrade from VuFind 1 to 2**.

A quick opinion survey among the attendees retrieves a greater number of those interested in an upgrade than those who already gained experience with it.

The question for a migration strategy leads to the estimation that a migration is not generally appropriate in case the crucial features are not applied at all. Also, if a big number of installations have to be serviced, which might partially be costumized very distinctly, a migration should not be carried out unconditionally. In any case, ressources for development are necessary, which add to the ongoing expense for maintenance. A strategy could be to implement newly set up installations with VuFind 2. The experience is reported that VuFind 2 pleasantly supports custumizations more neatly than version 1. Some experience with the Zend framework might be necessary ahead. For EBSCO customers there is the advantage of the easily arranged EBSCO update. Also, the concern is expressed that the community could concentrate increasingly, if not excludingly, on version 2 and new developments could cease to go on in parallel. Good reasons in favor of version 2 are the Bootstrap theme and responsive design, new translation mechanisms, upcoming hierarchical facets, and the utilization of different authentication mechanisms. However, the end-user features appear unchanged, apart from the upcoming hierarchical facets.

The discussion is followed by a session of **Lightning Talks**, which is presented again by Bettina Sunckel.

Markus Mächler (snowflake productions gmbh, Zürich) starts off with a presentation of Bootstrap 3 Theme. Bootstrap 3 Theme is offered since VuFind 2.3, and it will replace Bootstrap Theme with VuFind 2.4.
With Bootstrap 3 Theme, responsive design can be implemented with VuFind, that is, for different devices, the presentation of identical content can be determined to adapt to, e.g., the respective display geometry. In collaboration with swissbib, a redesign of the swissbib search is currently being implemented. In practice, using VuFind's current template structure, the content cannot be sufficienty divided up onto different files and the column layout is not consistent enough to fulfill the desired detailedness for swissbib's page reproduction. Therefore, adjustments of the core were necessary. A live demonstration nicely shows the obtained effect, especially the off-canvas navigation is very useful for smartphones, because the facets can be faded in when they are needed.

One subsequent question from the audience shows interest in adopting the modifications. This is possible because a good portion is already fed back into the core, though the off-canvas navigation is not yet completed.
Another question is directed at an integration of Typo3 and VuFind. However, this has not yet been taken into account at all.

In the second lightning talk, **Clemens Kynast** (ULB Jena) shows an approach to an alternative to the classic two-tab design of VuFind. Instead of either two columns or one integrated list, a solution has been implemented by adapting the Villanova-two-column theme to two tabs and one presented list at a time. The facets are arranged exclusively on the left und adapt accordingly to the presented result list.

A participant asks, whether this solution has already been tested with users. This will take place soon, in order to make sure that it is a useful modification.

As third lightning talk, **Martin Fuchs** (subkom GmbH), presents the employment of bib connectors with VuFind. The subkom GmbH offers its discovery system smartBib based on VuFind for the connection to existing and running library software and for the integration of external data sources. Especially for public libraries, smartBib is interesting because VuFind in itself is not an issue. The offered bib connectors are implemented in Java und can gather the relevant data for the respective VuFind functionalities using different modalities, like SQL queries or screen scraping.

The audience shows interest in which library systems smartBib can work with. These are, among others, Bibliotheca, BiBer, SISIS, aDIS/BMS, and concerto.
In addition, a question is asked concerning support of DAIA and PAIA. Support is possible using a connector.

In the 4th lightning talk, **Sven Stefani** (UB Kassel) reports about PUMA as publication management system. PUMA stores, device-independently, publication metadata and files associated with them, e.g., in pdf format, and manages the administration and organization of the publications via tags. The system is based on the BibSonomy engine and is open source. PUMA offers a REST-API and REST clients for Java, PHP, and Python, as well as browser add-ons. Publication data can be imported by scraping. Nearly all citation- and export formats are supported. The VuFind module allows for a substitution of the favorites list of VuFind by the data loaded from PUMA, and for the presentation with VuFind.

A complementary comment from the audience states that, after the login via VuFind, the employment of Shibboleth allows for a seamless transition to the PUMA system.

Oliver Schihin (swissbib), with the 5th and concluding lightning talk, goes into detail with jusbib, the meta-catalog of Swiss libraries of law and other relevant collections of legal literature (jus.swissbib.ch). Jusbib searches the swissbib index but uses a complicated but still well performing filter to make only legal literature visible. Additional information is accessible through other facets and subject headings. In particular, a Swiss law classification system has been integrated for jusbib, which is presented as hierarchical facets of multiple levels in the advanced search. Notations are translated. The development is ongoing work together with snowflake and done with partners like the Department of Justice, Swiss Supreme Court and the Association of Swiss Law Libraries. Together with other swissbib views the site will be migrated to a responsive design in 2015.

Afterwards, a question addresses the number of possible levels of hierarchy of the classification system. The answer states that the number is arbitrary and controlled by the points of the class notation. Classes and templates are public domain, but they will have to be adapted.

Till Kinstler (VGZ Göttingen) takes over the subsequent **session of “hints and clues” and "users ask - users answer"**.

The first topic addressed is **Shibboleth** and the advisability of setting up several service providers for different functionalities. Bettina Sunckel (HeBIS-IT) reports that for HDS (HeBIS Discovery System), only one service provider is used which works well. The access to online media, however, is a weak point, if publishers do not support Shibboleth throughout, and therefore for the HDS installation in Frankfurt, a proxy has been set up.
The difficult configuration of Shibboleth is mentioned in the discussion.
Also, the re-use of attributes of existing identity providers is discussed. Bettina Sunckel describes that in Frankfurt an existing connection to the user administration could be adopted for HDS.

The **de-duplication** of records is brought up as second topic. Nobody present seems to have already gained experience in using the record manager of the Finnish National Library.
One contribution to the discussion describes that fuzzy de-duplication in a matching process can yield quite useful results. Such a matching code will be published soon by the Cern library.
swissbib reports that they perform de-duplication with CBS.
One notice addresses the research project of the consortium of library networks (AG Verbundsysteme) which arrived at the conclusion that automatic de-duplication is not reliably practicable concerning their data on hand.

With regard to **Solr indexing**, a reproducible problem of one participant consists in a “broken pipe” error after 20,000 records which leads to the loss of index documents.
However, the underlying problem remains unclear, suggestions range from a timeout configured too tightly to some invalid record causing a complete abort. It does not seem to be related to the by now solved problem of Solrmarc with uncomplete batch import.

As 4th topic, the usage of **SolrCloud** and according experience is brought up. Till Kinstler reports that there is good experience. It is crucial to decide on the intended failure stability. One server might seem to be unreachable, e.g., because of the garbage collection being busy, and this must not occur on several servers at the same time.
An advantage of SolrCloud, even with a small index fitting into RAM, is the possibility to distribute the index across several machines and stability can be obtained by redundant data replication.

Bettina Sunckel thanks all participants and closes the meeting.