Appendix A Exploration of the do-it-yourself scenario

Content

1. Introduction

In this scenario, the national discovery tool for France would be newly developed: the metadata and possibly an index of the full text will be retrieved from each publisher. Users that are member of a library with a subscription to the resource will be given immediate access. For others, a delivery mechanism should be provided.

For this scenario, a number of case reports were studied: Trove (Australia), Suchkiste, Journals Online & Print service and the EZB linking service from Germany, the development of FINNA in Finland , the resource discovery programme of JISC in the UK and the situation in the Netherlands. These case studies are described in chapter 2 to 6.

The do-it-yourself scenario for the French national discovery tool is discussed in chapter 7.

2. The Netherlands

2.1 A common discovery portal for public and academic libraries has been studied

In the Netherlands, a co-operative effort of public and academic libraries focused on the national library infrastructure brought out a vision document that describes the objective to create one discovery service for the complete print and digital collections of the libraries in the Netherlands. As a first step, the possibilities to set up such a discovery service by combining the metadata of the digital content of the Dutch libraries were studied.

The University libraries decided not to participate in this potential new discovery service. Several years ago, two University libraries had setup their own locally developed discovery tools: the University of Utrecht with Omega en the Tilburg University with Get-It. The development of both discovery services took quite an effort. However, when a situation of steady state was achieved, the manpower involved in the management and maintenance of the discovery service was relatively low: one library estimated it at 1 hour per week, the other at 1 FTE. These discovery tools succeeded in acquiring metadata of 10-15 large publishers. However, the coverage of the locally developed discovery services was somewhat unsatisfactory: approximately an estimated 70% to 80% of the digital collection of the library was covered. The 'long tail' of scholarly publishers made it difficult for a smaller institution to have relationships with the smaller ones. At the moment, both University libraries have reconsidered their locally developed discovery services. Acquisition of a commercially available discovery service would bring three advantages: (1) higher coverage (2) benefiting from the continuous innovations by these providers and (3) compatibility with other management tools for the digital library such as link resolvers/knowledge bases and the library management systems in the clouds, presently in development. The University of Tilburg have migrated to OCLC WorldCat Local, the University of Utrecht has decided to forgo a web scale discovery system and focus instead of improving delivery mechanisms of standard search engines such as Google Scholar or bibliographic databasesf such as Web of Science.

2.2 New policy by the University libraries in development

The public libraries in the Netherlands are now developing a portal of their own. The plans of the University libraries in the Netherlands are still in development. A transition to a library management system in the cloud that will be shared by all Dutch universities is envisaged. This library management system should have the functions of shared cataloguing, shared acquisition and shared electronic resource management. It should be compatible with the various discovery tools that are already implemented the Dutch University libraries. At this moment, a common discovery service for all Dutch University libraries is not foreseen.

2.3 Information sources

- Interview Anja Smit and Marcel Rasch, University library of Utrecht
- Visiedocument GII Consortium, 2010
- Onderzoek naar de opties voor een centrale database met metadata van digitale content; Maurits van der Graaf; Pleiade Management en Consultancy; 2011.

3. The resource discovery program by JISC in the UK

3.1 Resource discovery task force vision and the JISC discovery program: a non-tool focus

The resource discovery task force vision focuses on facilitating the establishment of aggregations of (open) metadata, to which libraries, archives, museums and other resource providers can contribute. The vision and its subsequent activities of the JISC-funded discovery programme are documented at the website: http://discovery.ac.uk/ . The ultimate aim of the JISC discovery programme is to facilitate metadata aggregations in order to be used by discovery services rather than to develop a discovery service itself. The reason for this is that in an earlier stage the possibilities for the discovery programme were discussed with the wide range of librarians. From this consultation round, many different use cases were seen as relevant by the various representatives:

- There was a wide variation in proposals for the content the week offered by such a discovery tool: from scholarly literature and library content only to inclusion of digital archives and/or cultural heritage materials.
- In addition, there was a wide variation in user groups for the discovery tool: librarians, users in higher education and/or users in the cultural/heritage institutes.
- Different opinions and expectations about the interface: was it going to be to be a part of Google Scholar, a simple search tool, should it also function as a union catalogue etc.?

This variation of possible use cases was seen as too wide to be handled by *one* national discovery service. However, there was consensus about the importance of good quality metadata. Therefore, JICS set up the discovery programme with a non-tool focus, but focusing instead on facilitating aggregations of metadata in an open way, so that discovery services developed by others can use those metadata.

3.2 Knowledge Base + and the GOKb

Knowledge Base+ is a new shared service from JISC Collections that aims to help UK libraries manage their e-resources more efficiently. It is being established to start addressing the challenges facing libraries due to the inadequate data and metadata about publications, packages, subscriptions, entitlements and licenses that is available throughout the e-resource supply chain. Knowledge Base + focuses on data on JISC licensed content (NESLi2, SHEDL and WHEEL agreements)[1] and works with the ONIX-PL standard. It is important to note that it is not an electronic resource management system, but it focuses on the data and it can be used within an electronic resource management system[2].

[1] In the second phase, data will be added on more non-NESLi2 e-journals, full text databases, e-books and open access publications in order to make coverage as comprehensive as possible for UK libraries. Also, a project will be undertaken to gather more comprehensive information on institutional holdings and entitlements so that KB+ can be pre-populated with as much institutional data as possible.

[2] It is already proven that the data from Knowledge Base + can be loaded in the 360 resource manager from Serial Solutions (ONIX-PL, JISC collections and 360 resource manage; post on October 15, 2012 by Graham Stone).

As part of the Kuali Ole open source library management system development, the global open knowledgebase (GOKb) aims to become an open knowledgebase using standards-based architecture and with a CC0 license. The partners of Kuali Ole - over 20 American academic libraries - work together with JISC (Knowledgebase +) on this project. In the table below, the data elements that will be covered by the GOKb and the data elements that should be covered in the local ERM system of its library are presented[3].

Global Open Knowledgebase (GOKb)	**data elements**
community managed data	*title description*
accessible using API's	*standard ID*
Open (CC0 license)	*package (a.k.a. collection)*
way for libraries and vendors to share identifiers	*platform*
in local ERM system (e.g. Kuali Ole):	*subscription (deal)*
	purchase order
	issue entitlement
	license
	usage statistics

The aim is that the GOKb will interact with Knowledgebase + and other collectively managed knowledgebases.

3.3 Other issues

- **COPAC :** the union catalogue by the RLUK (a consortium of over 20 research universities) is at a moment of widely used and seen as an aggregation of metadata that can have multiple uses: for example, not only for search purposes, but also for collection management services. Therefore, there are no plans to replace COPAC by a discovery service.
- **Shared library management systems:** JISC also has a programme for the development of shared library management systems. However, in the thinking by JISC, a library management system is focused on the printed collection and does not necessarily include an electronic resource management function. In this view, library management systems are squeezed in the middle (by electronic resource management systems on the one hand and discovery services on the other hand).
- **Data formats:** The interviewee expresses interest in the choices that ABES might make with regard to the data formats that will be used by the national discovery tool. According to this interviewee, the MARC data format is at the end of its life cycle. Newer data formats are JSON, MODS and linked data (RDF). The Library of Congress has announced that they will migrate from MARC data format towards linked data. It is noted that linked data are very expensive to produce. The JSON data format has the advantage of easy visualization. The interviewee states that he does not know which way it will go and if there will be one winner or that in the future discovery systems have to cope with different data formats.

[3] From: Introducing the Global Open Knowledgebase (GOKb), Maria Collins and Kirsten Wilson, NCSU libraries; PowerPoint at ER&L 2012

3.4 Information sources

- Interview Andrew McGregor, JISC; interview Liam Earney
- JISC and RLUK; One too many; many to one: the resource discovery task force vision

4. Trove by the National Library of Australia

4.1 Development and launch of the discovery service Trove

In September 2008 the National Library of Australia started a project to develop a new discovery service. The new service was released in December 2009 under the name Trove. Trove replaced eight legacy discovery services (including the Australian National Bibliographic Database), and aimed to improve the discovery experience for the Australian public and researchers by including more content and by allowing users to engage with the content. The NLA chose to undertake this project as an in-house development, rather than using a vendor's product because of the (at that time) innovative character of the discovery service. Trove covers among others newspapers, pictures/photos, music/sound and video and maps.

The development effort took one year and four months. The development team consisted of 2 developers, 1 user interface designer, 1 business analyst and a project manager. The total effort is roughly estimated in terms of manpower at less than 10 person-years, in terms of money at over AU$500,000. Since then, the further development is an iterative process. The same team is involved in the maintenance of Trove but is also involved in other projects.

4.2 Effort to increase coverage of journal articles

Trove covers also scientific literature. In Trove Stage 4 (the development stage running from 2010 to 2011), efforts were made to increase and improve the coverage of (digital) articles of scholarly journals. One part of the efforts was focused on covering more article metadata, another part on improving the access to journal articles. This additional effort lasted approximately 5 months and is estimated to have taken approximately 3 person-years (an estimated AU$340,000).

With regard to the effort to offer more journal article metadata, the effort has resulted in covering approximately 250,000 journal articles. An important and time-consuming problem in covering more metadata from various providers consisted of the different, non-standard data formats and the continuing changes in those data formats by the providers. Also, some providers were not willing to deliver article metadata.

With regard to the effort to provide access to journal articles, a user authentication system had to be set up as well as databases with license and holding data of the various libraries in Australia and with databases to enable the authentication system[4].

This has resulted in the following authentication mechanism in Trove (see also figure 1):

- When viewing the article metadata, a link to the library/online holdings of the journal is shown (based on the ISSN and or ISBN or another journal identifier). It is also attempted to show only the libraries which have the relevant issue of the journal.

[4] A database of all Australian library EZProxy server addresses and local IP address ranges; a database of "short library names", to help Trove users recognize and select their library by name; for all libraries without EZProxy servers, a database of Australian library login web addresses and associated information, mappings from Trove library codes to the vendor library codes.

- For online e-resource articles, the libraries that have this particular article in the holdings will be shown.
- Users can be identified via IP address or via a registration procedure (library membership).
- Access can be provided in the following ways:

- View online[5]:
- users affiliated with a library with a proxy-server are referred to the proxy which will then pass them on to the article.
- users affiliated with a library without any authentication mechanism known by Trove, the user will be linked to the article on the vendor sites: the vendor is then responsible for the authentication of the user. The same mechanism applies for the users without library affiliation.

- Borrow/Buy:
- there is also a window with the option borrow, which list the libraries that hold the journal
- there is also a window with the option buy, which links to the document delivery service of the National library of Australia.

4.3 Effects of Trove and lessons learned

Trove attracts approximately 50,000 visitors per day. The newspapers are the top attraction: 85% of the usage is focused on those. Trove is used by the general public (such as family historians), but also by researchers at the Australian universities. A number of further developments are described in the strategic plan:

- An API is in development that will allow other discovery services to use Trove as a target.
- A number of efforts are undertaken to increase (1) the coverage of Trove, (2) enhance its usage and (3) develop communities of contributors and partners.

With regard to the scholarly journal literature, the present situation is expected to be maintained. The need to cover more scholarly journal literature is seen as not urgent, since a number of the larger Australian universities have implemented webscale discovery services such as Summon, WorldCat Local, EBSCO Discovery or Primo. Providing a more comprehensive coverage will be too costly for the NLA. In addition, in developing stage 4, the development team encountered the following problems that were difficult to solve:

- The conversion to a unified data format of the different, non-standard data formats and the continuing changes in those data formats by the various providers was time-consuming.
- Some providers/publishers appeared not willing to deliver article metadata.

[5] Two processes mentioned in figure 1 are not implemented as access was already provided to all articles via the other methods. These processes are:

- users affiliated with a library with an OpenURL resolver are referred to article at the vendor site via the link resolver.
- users affiliated with a library where screenscrape authentication has been implemented, Trove will link the user to a Trove page requesting users' library login details. Trove will then authorise the use and retrieve the article or inform the vendor about the authorisation.

- It appears to be rather difficult to keep the information on the subscriptions and licences by the various libraries up to date. In addition, some libraries subscribe to customised collections.

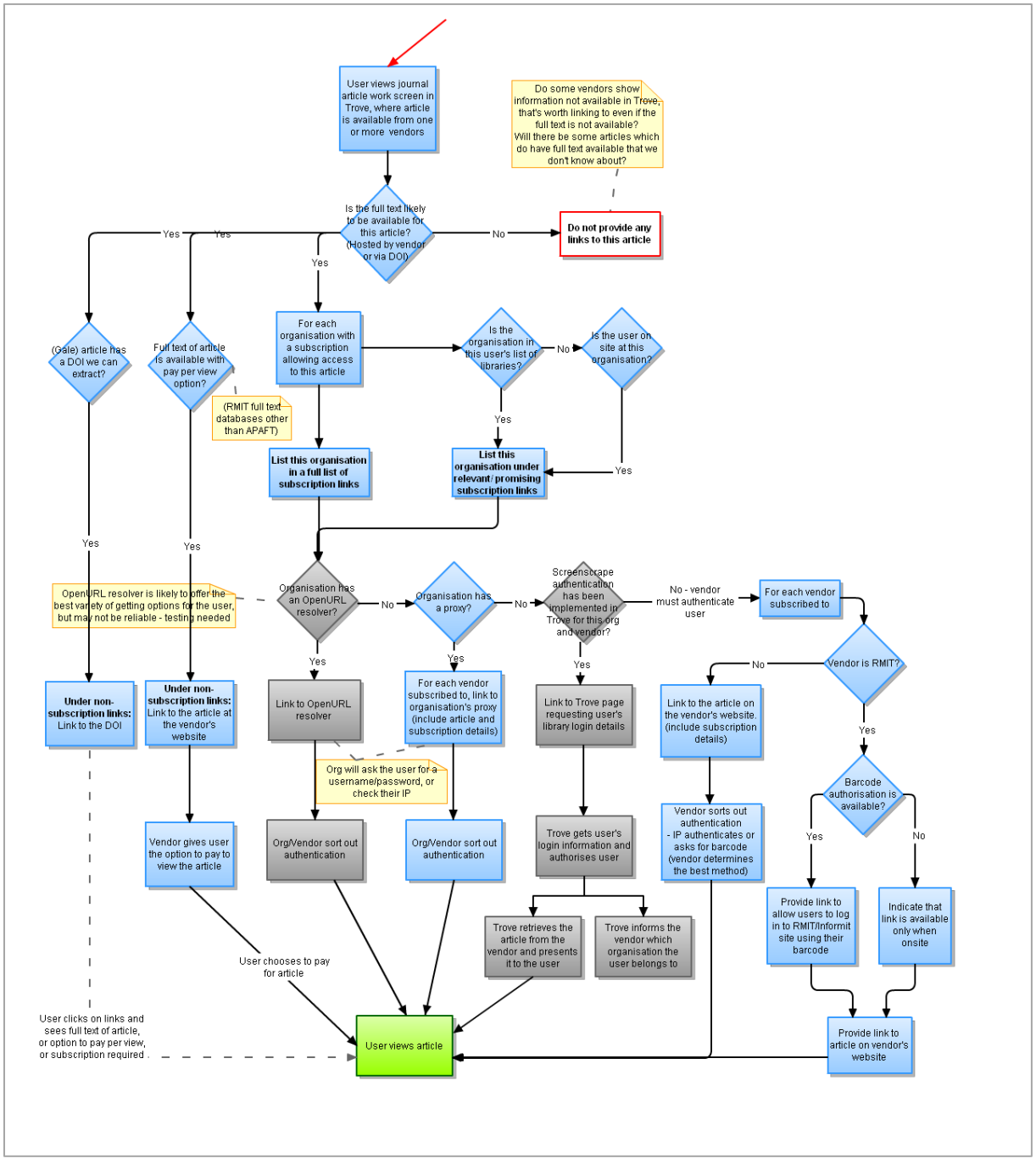

Figure 1 Viewing a journal article in Trove

4.4 Information sources

- Interview with Mrs. Susan Collier, director, Collections Access Branch, IT division, National Library of Australia
- USER AUTHENTICATION FOR E-RESOURCES WHICH WILL BE ACCESSED VIA TROVE: A DRAFT MODEL; Working Draft: 2 December 2009; NSLA Open Borders project.
- Developing Trove: the policy and technical challenges; February 2010; Warwick Cathor, Susan Collier.

- Trove Stage 4 - Journal articles and e-resources; 1 November 2010
- Strategic plan, July 2010 to June 2012, National Library of Australia

5. The national library infrastructure in Germany in relation to discovery

5.1 ZDB and EZB

Two important library services with regard to scholarly journals within Germany are ZDB and EZB**:**

- ZDB (Zeitschriftendatenbank) is a union catalogue for integrating resources (print- and e-journals, newspapers, e-papers, serials, etc.). ZDB contains 1.5 million bibliographic records and 11.8 million holding and license information records of 4300 German and Austrian libraries. ZDB also imports license information for e-journals from the Electronic Journals Library (EZB). ZDB is a service of the German National Library and the Staatsbibliothek zu Berlin.
- EZB (Elektronische Zeitschriftenbibliothek) is a standardised platform with bibliographic information on digital scholarly journals. In 2010 it covered over 52,000 journal titles, approximately half licensed journals and the other half Open Access journals. EZB is used by 555 German libraries and over 100 libraries in other countries (43 in Austria, 27 in the Czech Republic, 16 in Slovakia and 19 in Switzerland and a few in other countries). EZB is a service of the University Library of Regensburg.

5.2 A closer look at the EZB services

- **Efficient workflow for libraries participating in the EZB:** The EZB is maintained by a collaborative effort of over 600 libraries. The EZB contains bibliographic data of electronic journals (title level) and license data and holding information for each member library. Member libraries can update their licence data via easy-to-use webforms. In addition, data on national licenses and data on licenses bought by consortia are also added to the EZB with special functionality so that this information is available for the libraries that are involved with these licenses. The effect of this is that member libraries use the EZB to maintain and update their holdings information with regard to electronic journals and download this information from the EZB to their own local catalogues and to the knowledgebase of their link resolver. At this moment, EZB is working on an interface to facilitate downloading to the knowledge bases of webscale discovery services (these have slightly different formats than the knowledge bases of link resolvers) and will use the KBART standards to achieve this.
- **Search options for end-users**: end-users can use the EZB for searching journal titles. Another important end-user service is the EZB linking service (see below). The EZB has plans to set-up an article search service in the future. However, no concrete actions are taken yet because of the expectation that collecting and processing the article metadata will be very labour-intensive.
- **The EZB linking service:** The EZB linking service is based on the Open URL technology and includes all e-journals in the EZB. The EZB link resolver is integrated in over 40 information

services, such as the EconBiz search portal and Medline. The EZB Linking service offers direct article linking for over 20,000 e-journals. The EZB link resolver can be used in the following ways:

- as an independent link resolver service in order to link to electronic full text in digital journals
- in connection to a link resolver by local library via two methods:
- The licence information (with indications of the time periods and the URL's for the linking to the full text) can be loaded into the knowledgebase of the local link resolver[6].
- The local link resolver can use the EZB link resolver as a target in itself.
- **Manpower involved**: as mentioned earlier, the content of the EZB is maintained and updated by a collaborative effort of the over 600 member libraries. The manpower management and further development of the technical infrastructure (including the EZB link resolver) is estimated at 1 to 1.5 FTE for IT staff members (of which an estimated 0.5 FTE for the link resolver).

5.3 JOP - Journals online & print

Based on ZDB and EZB, the service Journals Online & Print (JOP) aggregates the holdings information for participating libraries about the journals and journal collection from ZBD and EZB. This means that a library can get easily the information about its own collection in an integrated way. The service delivers uniform data on licenses and data on printed and electronic journals. The JOP could be described as a knowledgebase, although a limited one. The 'knowledgebase' of JOP contains information about several hundreds of packages of providers, whereas commercial knowledgebases by for instance Ex Libris or OCLC contain thousands of such packages. The management of the knowledgebase of JOP is done manually involving a co-operative effort of more than 30 librarians (including cataloguing). This is a lot of work and does not scale. Therefore, a project has been started to make an automatic updating from providers possible.

The JOP has an OpenURL-based web service that in combination with a database will provide the end-user of a specific library with information[7] indicating if the journal is available in print or online. If an online journal is available context sensitive links to the journal or full text are provided. In effect, this web service of JOP can have a similar function as a link resolver for smaller libraries without a link resolver of their own.

For larger libraries with their own link resolver, the JOP web service is hardly relevant. These libraries could use the information about their own collection from JOP and put that in the knowledgebase of the link resolver. However, this is not (automatically) possible because of the fact that in Germany, a special (and very good) journal identifier (the ZBDID) is used, but often not recognised by the international commercial knowledgebases.

The respondent finds it difficult to give an exact estimate of the effort involved in the development of JOP as it was part of a larger project. A rough estimate is less than one person year for software design, development, testing and documentation. An ontology for availability information is available at <http://www.gbv.de/wikis/cls/Document_Availability_Information_API_%28DAIA%29> & <http://www.gbv.de/wikis/cls/DAIA_-_Document_Availability_Information_API>. If there are

[6] The EZB link resolver gives access to over 20,000 journal titles. Many (larger) academic libraries have a link resolver of their own in order to give access to other electronic publications as well.

[7] There are 3 different APIs for this service: Icon, XML, HTML

structured APIs (like SRU or PSI) available for the data involved and a modern web framework is used, the development of such a web service would take much less effort than one person year. The respondent estimates that one could develop a prototype within weeks.

5.4 Suchkiste: a discovery service for nationally licensed content

Suchkiste started as a project to develop a discovery interface for the journals in the EZB-service that fall under the national licenses. The project was originally funded by DFG and has been carried out by the University of Göttingen. The resulting discovery service (http://finden.nationallizenzen.de/) made use of VuFind for the interface and Solr for the search engine - both open source software.

Suchkiste might serve as an example for the do-it-yourself scenario for building a national discovery tool in France. The following facts might therefore be relevant:

- **Authentication**: via Shibboleth or EZProxy[8]
- **Metadata and national licenses**: it is described that the effort to make the metadata uniform and enrich with specific German identifiers for journals was very labour-intensive. The metadata were coming from over 40 different providers with many different formats and quality standards.
- **National licenses**: to date, Germany has invested approximately hundred million euro in the national licenses. The target groups are scientists, students and scientifically interested private persons. Especially the last target group is outside the scientific library domain and for them a special authentication system had to be set-up. Registration is required: approximately 8000 private persons are active users of Suchkiste.
- **Solr:** for the index platform, the open source software Solr is used because of its high performance. Primo is using Solr as well for its index. The index in Solr produced by Suchkiste is also integrated in Primo via the Solr interface. This means that Primo clients in Germany can use this index.
- **Result ranking**: institutes that are registered for the national licenses get a special URL: this is recognised by the discovery tool and this results in a results ranking that is adapted to the own choices of relevant content.
- **Opening up for internet search engines**: the index of Suchkiste is also made available to Internet search engines like Google or Google Scholar, so that the URL of the national license is shown in the search results.

The lessons learned by Suchkiste are the following:

- **Integration in discovery tools and link resolvers:** Libraries have the choice of receiving the bibliographic metadata or use the index of the Suchkiste directly. At the start the majority of the libraries preferred receiving the metadata; now the majority is using the index (via Solr sharing) in their own discovery tools. Link resolvers are not necessary, since all libraries can use the same URL to link directly to free articles on the publishers. To achieve this, however, a complex knowledgebase and (proprietary) ERM are necessary.

[8] For this, more than 72000 IP address entries were manually registered of over 670 institutes. This is a rather error prone effort.

- **IT**: For technical maintenance and operation a constant 4 hours per month is needed. The system itself does not have its own hardware, but is part of the general infrastructure of VZG.
- **Uploading new metadata**: The most labour-intensive part of the maintenance is the processing of the metadata. The manpower needed for this can vary from a few hours for a certain package to a few weeks. The costs for this processing are financed by the buyer. Regularly, providers sent undocumented data dumps in proprietary formats: labour-intensive reverse engineering might be necessary in those cases.
- **Usage**: The entire system can be harvested by Internet search engines. About 1300 people per day use the Suchkiste and 95% of these users come from hits on Google.

5.5 Future plans for the library infrastructure in Germany

The Deutsche Forschungsgemeinschaft proposes service structuring process that moves away from regionally organised structures and towards functional and nationwide oriented services. DFG will initiate and support the restructuring process. For this purpose, the call for proposals focusing on four areas for which the development of new structures and services is seen as most urgent. Two are relevant for this study:

- **The library data infrastructure and local systems:** the aim is to promote a functional uniform cataloguing and data platform. The shared system would give to basis for searching, availability and administration of printed and digital library databases. Now these data are in different silos. One mentions monographs, print and electronic media and other types of data/content.
- **Electronic resource management**: this focuses on the development of components for a nationally available, shared electronic resource management system with the main objective to create the possibility of a uniform nationwide use of the data for managing licenses on the local, regional and national level.

5.6 Information sources

- Interview Evelinde Hutzler - Universitätsbibliothek Regensburg, Elektronische Zeitschriftenbibliothek
- Interview Johan Rolschewki - Staatsbibliothek zu Berlin
- Interview Sigrun Eckelmann - Deutsche Forschungsgemeinschaft
- Interview Gerald Steilen - Verbundzentrale des GBV (VZG)
- Taking digital transformation to the next level; the contribution of the DFG to an innovative information infrastructure for research; July 2012; Deutsche Forschungsgemeinschaft
- Auscchreibung ‘Neuausrichtung überregionaler informationservices”; 15.10.2012; Deutsche Forschungsgemeinschaft
- Suchkiste; DFG-Projekt der VZG; 15 Februari 2011
- JOP und Co; presentation Dr. E. Hutzler; J Rolschewski; Dt. Bibliothekartag 2009, Erfurt
- Journals online & print; brochure
- Electronic Journals Library; annual report 2010
- Der Schnelle Weg zum Volltext – Einsatz und Nutzung des Verlinkungdienstes der Elektronische Zeitschriftenbibliothek; E. Hutzle, M. Scheuplein, P Schröder; Bibliothekdienst 40 (2006), H. 3 p 306-313

6. FINNA - the National Digital Library of Finland

6.1 Development and architecture

In the document National Digital Library (2011) 'the creation of a joint public interface for the materials and services of libraries, archives and museums' is mentioned as one of the main objectives. The public interface will make searches possible for end users in restricted and unrestricted sources and in the long-term preservation system.

The project for the public interface started with the development of the specifications in 2009. After that, a tender was called. In 2010, ExLibris was selected with Primo. However, after a pilot project in 2011, it was decided that the public interface of the ExLibris software was not fulfilling the specifications set by the National Digital Library. Therefore, in the beginning of 2012 VuFind open source software was selected to develop the public interface. This public interface will be (among other indexes) connected to the Primo Central index of ExLibris.

The demo version of the public interface is: http://vufind-fe-kktest.lib.helsinki.fi/institution/

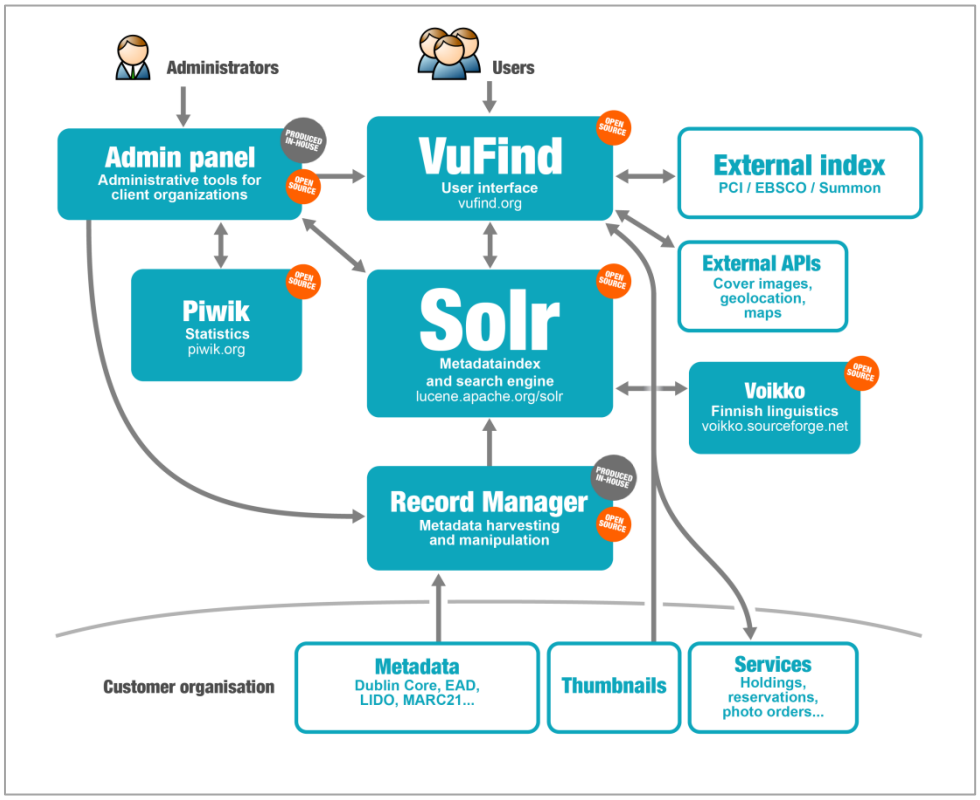

Figure 2 Overview public interface architecture National Digital Library

In figure 2 an overview of the public interface architecture of the National Digital Library is presented. It contains the following elements:

- The Open Source VuFind library resource portal.
- A Solr metadata index and search engine platform with the index of the Finnish cultural heritage content.

- An external index: this is the Primo Central Index with a large index of the scholarly literature maintained by ExLibris (EBSCO and Summon are listed in the picture as they were also tendering at the time of the drawing of this picture).
- A Record Manager: a component that is involved in metadata harvesting and manipulation (of the Finnish cultural and heritage content) and could be compared with regard to it solution with the Metadata Hub of ABES. This is in-house developed but will be open source.
- An Open Source component with regard to the Finnish language that presents the end-users with spelling suggestions. This component already existed.
- Piwik: an open source program to collect user statistics.
- An Admin module providing administrative tools for participating organisations. This is in-house developed that will be open source.

The FINNA development is focused on using open source packages in combination with commercial software packages. With regard to commercial packages, FINNA makes use of the above-mentioned Primo Central Index covering a large part of the international scholarly literature, of the SFX link resolver and of the BX recommender service – all from ExLibris. MetaLib - the federated search engine by ExLibris - is also used by many Finnish Higher Education institutes and connected to the present FINNA set up, but it is envisaged to be replaced by the Primo Central Index.
Finnish academic libraries use a central electronic resource management system for nationwide licenses. Nationwide licenses are estimated to cover approximately 80% of the digital collection of higher education institutes. The other 20% of the digital collection is managed by local management systems. This nationwide ERM system feeds into the knowledgebase of the SFX link resolver. There are plans to develop this ERM system further in such a way that also individual licenses can be administrated.

6.2 Portal *and* gateway

FINNA aims to become the portal *and* gateway for end-users for Finnish libraries, archives and museums. At the moment of writing, FINNA covers about 10 organisations but it will expand to over 400 organisations in the coming years. Most of those organisations fall under the Ministry of Education and Culture. The gateway function of FINNA will directly connect the end-user to the backend systems of those organisations in order to see if a certain item is available, reserve it for borrowing, pay for it etc. In the longer-term, FINNA is expected to replace the local web interfaces of the institutes involved, thereby increasing efficiency and cost effectiveness. Organisations that participate in FINNA are asked to sign a service contract about their responsibilities etc. For organisations falling under the Ministry of Education and Culture, no financial costs are involved to participate in FINNA. However, for other organisations a fee might be involved, but this is not yet developed as this will come at a later stage.

Other aspects of FINNA are:

- FINNA has a connection to FeedNavigator, a service that shows recent articles on the same topic

- FINNA will be connected to the union catalogues Helka and LINDA[9], but this not a major aim as union catalogues do not have availability information, which is seen as part of the gateway function (see below).
- The connectors (API's) of the portal software VuFind with the backend systems of the organisations are seen as crucial because of the gateway function. VuFind is particularly accommodating for these connectors. At the moment a number of catalogues are connected to FINNA, making it possible for an end-user to see the availability of a certain item and to reserve it (after authentication as a member of the library).
- The policy is to open up the metadata of Finnish cultural and heritage content that are established and indexed for discovery by FINNA for other discovery services and search engines by a CC license. However, this is not yet fully implemented as some contracts with suppliers of metadata prohibit this at this moment.
- The authentication mechanism for the HE institutes works via Haka - a Federated authentication system implemented at all Finnish HE institutes.

With regard to the manpower needed for development and maintenance of FINNA the interviewee indicates that 5 FTE is involved in the technical development and maintenance of the system, while there are 12 FTE involved in communication and training of the over 400 institutes involved. Especially smaller institutes without IT staff require much attention. A very rough estimate for the development of the present public interface is about 10 to 15 person-years. An additional rough estimate for the investment needed for the hardware is 100.000 to 200.000 euro.

6.3 Information sources

- Interview Kristiina Hormia-Poutanen, Deputy National Librarian, Director, National Library Network Services National Library of Finland
- National digital library-enterprise architecture; www.kdk2011.fi
- Libraries, archives and museums working together!; Presentation LIBER 2010; Kristiina Hormia-Poutanen
- The National Digital Library - collaborating and interoperating; Ministry of Education and Culture 2011;26
- Public interface functional requirements; specification 1.1 draft (11 September 2009)
- National Metadata Repository Project; http://www.nationallibrary.fi/libraries/projects/metadatarepository.html

[9] HELKA is the joint database of the Helsinki University Libraries and the National Library of Finland. HELKA contains information about acquired books, periodicals and serials. You can search for materials in the HELKA online catalog. LINDA is the union catalogue of the Finnish University Libraries, also including the National Repository Library, the Library of Parliament, the Library of Statistics and Lahti Science Library.

7. Discussion of the do-it-yourself scenario

7.1 Components of a national discovery system

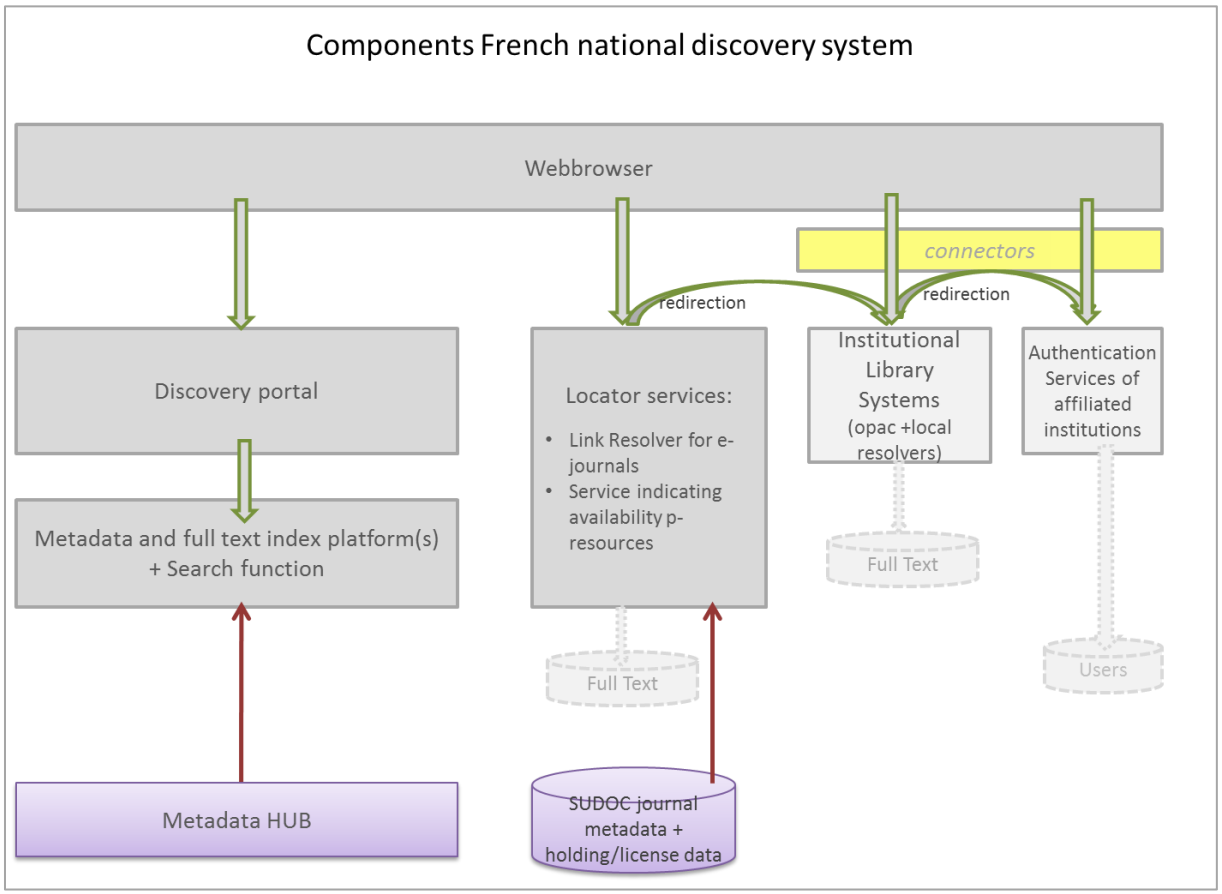

Figure 3 Components national discovery system

From the exploration of the do-it-yourself scenario, the following main components of the national discovery service can be identified (see also figure 3):

- **A discovery portal**: the portal presents the user interface and provides the connections with the other components. The requirements 10 to 18 are directly relevant for this component (see for the requirements chapter 2 of the main report).
- **A metadata and full text index platform and search function**: the portal connects to a platform with metadata and/or full text indexes of the scholarly literature, also called a centralised index. Requirement 8 and 9 are especially relevant for this component. In addition, the Metadata Hub of ABES that will enrich metadata will feed into this platform (red arrow in figure 2; requirement 7).
- **Locator services (link resolver and webservice indicating availability)**: locator services - a link resolver for electronic journal articles and a web service indicating availability for printed resources – will point end-users to access of full text provided by their library (either digital or print collections). Requirements 1 to 6 are all relevant for this component. The locator services will have to use a knowledgebase with information about various collections of the French HE libraries. SUDOC already contains an important part of the information needed (red arrow in figure 2; requirement 3).

- **Connectors to institutional systems (OPAC and authentication services):** after discovery of a print item, the end-user will be connected/redirected to the OPAC of their library to see if the item is directly available. A further connection (redirect) to the institutional authentication service will be needed to enable the user to reserve this particular item (see also requirement 4).

7.2 Portal

With regard to the interface, the open source software VuFind appears to be a logical candidate to use if ABES were to decide to build the national discovery tool itself. VuFind is used by Suchkiste and by FINNA and is maintained and further developed by a collaborative effort of a number of academic libraries (see www.vufind.org). The FINNA interface was used as a basis to check the various requirements with regard to the interface (see chapter 2 of the main report for a listing and explanation of the requirements). The results are listed in Appendix C. Clearly, most requirements are met by VuFind.

7.2 (Meta)data platform

7.2.1 Contents covered (requirement 9)

In chapter 4 is described how an extra effort was made to cover more scholarly content by Trove, resulting in approximately 250,000 journal articles - a small part of the total journal literature**10**. In the experience of Trove, getting the metadata took quite an effort while processing the metadata proved to be labour-intensive. The interviewee described that publishers sometimes change their metadata formats at short notice, again requiring manpower for processing.

Discovery tools set up by individual university libraries in Utrecht and Tilburg (the Netherlands) succeeded in acquiring metadata of approximately 10 to 15 large publishers, but additional efforts to increase the coverage of the discovery tool were hampered by the so-called long tail of scholarly publishers. A plan by EZB to setup a discovery service at the article level has been postponed because of the expectation that the acquisition and processing of the metadata would require too much manpower. In addition, the interviewee from Suchkiste described that the processing of the metadata for the nationally licensed journals was very labour-intensive.

7.2.2 Levels of metadata indexed (requirement 8)

Trove and Suchkiste produce their own indexes with metadata for scholarly journal articles. Full text indexing is not part of their indexes. Citation links are also not included. FINNA uses the Primo Central index of ExLibris for the scholarly content and thus follows their specifications. VuFind (and thus Suchkiste and FINNA) use Solr indexing and search engine for their systems (Trove uses Lucene; Solr is an extension of Lucene[11]).

7.2.3 Metadata enrichment and redistribution (requirement 7)

As Trove made its own indexes, metadata enrichment and redistribution appears not to be an issue. The interviewee from Suchkiste particularly mentioned the labour intensive processing of the normalising of metadata after delivery by publishers and other providers. Redistribution apparently

[10] Annually, an estimated 1.5 to 2 million journal articles are published in peer-reviewed journals.

[11] http://lucene.apache.org/solr/features.html

was not seen as an issue as well (for example: the resulting index was shared as well as the metadata).

FINNA also wants to make the metadata that has been collected and indexed by itself available to other discovery systems and/or search engines. However, there are still some issues with regard to copyright issues on some metadata to be solved before this policy can be implemented.

It appears that when the index of metadata is acquired and processed by the organisation itself, there is generally quite an amount of freedom with regard to processing, de-duplication, enrichment and redistribution because of the direct relations (and agreements) with the original producer of the metadata.

7.2.4 Sharing of a data platform

An example of interoperability of a national discovery service with discovery services of local libraries **(requirement 1)** is presented by **Suchkiste**. This is a discovery service for the journals in the EZB-service (see 3.3.1) that fall under the national licenses of Germany. Suchkiste makes use of VuFind for the interface and Solr for the search engine - both open source software. Primo is using Solr as well for its index. The index in Solr produced by Suchkiste is also integrated in Primo via the Solr interface. This means that Primo clients in Germany can use this index. Libraries have the choice of receiving the bibliographic metadata or use the index of the Suchkiste directly. At the start the majority of the libraries preferred receiving the metadata; now the majority is using the index (via Solr sharing) in their own discovery tools. Link resolvers are not necessary, since all libraries can use the same URL to link directly to free articles on the publishers. To achieve this, however, a complex knowledgebase and (proprietary) ERM are necessary. As Suchkiste is focused on nationally licensed content, this is also an example for **requirement 6** (Interoperability with a platform with nationally licensed content).

7.3 Locator services

7.3.1 A national link resolver

The **EZB linking service** in Germany provides an example to integrate a central system (EZB) with link resolvers from local libraries **(requirement 2)** and can be seen as a national link resolver. The EZB (Elektronische Zeitschriftenbibliothek) is a standardised platform with bibliographic information on digital scholarly journals. The EZB linking service is based on the Open URL technology and includes direct article linking for many e-journals in the EZB. The EZB link resolver is integrated in over 40 information services. The EZB link resolver can be used in the following ways:

- as an independent link resolver service in order to link to electronic full text in digital journals
- in connection to a link resolver by local library via two methods:
    - The licence information (with indications of the time periods and the URL's for the linking to the full text) can be loaded into the knowledgebase of the local link resolver.
    - The local link resolver can use the EZB link resolver as a target in itself.

In addition, the EZB is developing an interface for its database to interact with knowledge bases using the KBART standards. This can be seen as an important step towards interoperability with shared library management systems in the cloud **(requirement 5)**.

Another example is provided by **Trove**. Trove has set-up an authentication mechanism to provide access to end-users of different libraries, including libraries with link resolvers. See for a description below.

7.3.2 Locator services for libraries without link resolvers and/or for p-resources

An example to support providing access to content for end-users of libraries without link resolvers **(requirement 3)** is shown by the service **Journals Online & Print (JOP)** in Germany. Based on the union catalogues ZDB and EZB, JOP aggregates the holdings information for participating libraries about the journals and journal collection from both catalogues. This means that a library easily can get the information about its own collection in an integrated way. The service delivers uniform data on licenses and data on printed and electronic journals. The JOP could be described as a knowledgebase, although a limited one. The 'knowledgebase' of JOP contains information about several hundreds of packages of providers[12]. The management of the knowledgebase of JOP is done manually involving a co-operative effort of more than 30 librarians (including cataloguing). The JOP has an OpenURL-based web service that in combination with a database will provide the end-user of a specific library with information [13] indicating if the journal is available in print or online. If an online journal is available context sensitive links to the journal or full text are provided. In effect, this web service of JOP can have a similar function as a link resolver for smaller libraries without a link resolver of their own.

Trove provides another example of the inclusion of a union catalogue (the Australian national bibliographic database) so that when viewing the article metadata, a link to the library/online holdings of the journal is shown (based on the ISSN and or ISBN or another journal identifier). This set-up could be seen as a combination of requirement 2 and 3. Trove shows only the libraries which have the relevant issue of the journal. For online e-resource articles, the libraries that have this particular article in the holdings will be shown. In addition, Trove supports end-users in getting access to the full text of articles as much as possible. A user authentication system had to be set up as well as databases with license and holding data of the various libraries in Australia and with databases to enable the authentication system[14]. Access is provided as follows:

- Users can be identified via IP address or via a registration procedure (library membership).
- Access can be provided in the following ways:
    - View online:
        - Users affiliated with a library with an OpenURL resolver are referred to article at the vendor site via the link resolver.
        - Users affiliated with a library with a proxy-server are referred to the proxy that will then pass them on to the article.
        - Users affiliated with a library where screenscrape authentication has been implemented, Trove will link the user to a Trove page requesting users' library

[12] Compared to commercial knowledgebases this is rather limited: these contain thousands of such packages.

[13] There are 3 different APIs for this service: Icon, XML, HTML

[14] A database of all Australian library EZProxy server addresses and local IP address ranges; a database of "short library names", to help Trove users recognize and select their library by name; for all libraries without EZProxy servers, a database of Australian library login web addresses and associated information, mappings from Trove library codes to the vendor library codes.

login details. Trove will then authorise the use and retrieve the article or inform the vendor about the authorisation.

        - Users affiliated with a library without any authentication mechanism known by Trove, the user will be linked to the article on the vendor sites: the vendor is then responsible for the authentication of the user. The same mechanism applies for the users without library affiliation.
    - Borrow/Buy:
        - There is also a window with the option borrow, which lists the libraries that hold the journal
        - There is also a window with the option buy, which links to the document delivery service of the National Library of Australia.

7.4 Connectors

The example of integration with OPACs is provided by FINNA. In this discovery service (still in development) the union catalogues Helka and LINDA will be integrated as well as OPACs of a number of public libraries. Per item the availability can be shown and in some cases the user can reserve the item after logging in with his/her username and password directly from the user interface from FINNA[15].

7.5 Other functionality requirements

- Open API platform: the VuFind interface used by FINNA and Suchkiste has a number of APIs to interact with the search.
- User statistics: VuFind has the option to collect user statistics. FINNA has included an open source component (Piwik) to collect statistics.

[15] The VuFind software has an option to connect using AJAX protocol querying the catalogue.

7.6 Discussion

7.6.1 Match with requirements

No.	Overview of the results of the study into the do-it-yourself scenario with regard to the requirements for the national discovery tool	
1	Sharing (parts of) the index or metadata with other discovery services	Metadata and/or Solr index sharing – example Suchkiste
2	Interoperability with local link resolvers/knowledge bases	Example of EZB link resolver and interface with knowledgebases of link resolvers and discovery services
3	Interoperability with union catalogue in order to give availability information	Examples by Journals Online and Print and Trove
4	Integration/interoperability with local OPACs and ILL service	Examples by FINNA and Trove
5	Interoperability with the knowledgebase of the future shared library management system	Example of EZB interface with knowledgebases using KBART standard (in development)
6	Interoperability with a platform with nationally licensed content	Example of Suchkiste
7	Options to deduplicate, enrich and redistribute metadata	See paragraph 4.2 (Suchkiste; EZB)
8	Requirements with regard to the metadata and/or full text indexed	Indexing of full text and citation links not observed
9	Requirements with regard to the coverage of the scholarly content and the option to add 'private' content to the index	Extended coverage of scholarly content seen as (very) labour intensive and difficult to achieve
10	Search options	VuFind appears to offer the majority of the desired options
11	Non-English language support (as part of search options)	FINNA using the VOIKKO language support package for the Finnish language
12	Recommender options	See example by FINNA
13	Presentation of the results	See example by FINNA
14	Export options	Not observed
15	Sorting options	See example by FINNA
16	User accounts	See example by FINNA
17	Social features	See example by FINNA
18	Open API platform; opening-up mechanism metadata for internet search engines	VuFind offers API; Suchkiste has example of opening up index to internet search engines
19	User statistics	VuFind offers user statistics

Table 1 Match of the requirements for the do-it-yourself scenario

Based on the data collected in this study, an overview of the matching of the requirements for the do-it-yourself scenario is presented in table 5. The most important requirements that are not met in the discovery services studied are:

- **Data platform**:
    - The discovery services that built their own indexes have a limited coverage of the scholarly journal literature as many noted the difficulties to get metadata from (a large

number of) publishers and providers as well as the labour-intensiveness of processing those metadata in order to fit them in the normalised scheme used by the index of that particular discovery service (**requirement 9)**. The notable exception is FINNA, who uses next to its own indexes the index of Primo Central Index from ExLibris for its coverage of the scholarly literature (which is partly full text indexed).

- Full text indexing is not reported by the discovery tools studied here as well as the inclusion of citation links (**requirement 8).** With regard to full text indexing, FINNA is the exception as they make use of the Primo Central Index.

- **Portal:**
  - The specific non-English language support functions that are considered desirable to support the French language were not observed in the studied discovery tools. FINNA had included this feature, but this is delivered by a special Open Source language support package VOIKKO for the Finnish language. It is not known if a similar software package exists for the French language. However, Economists Online (a portal for economics literature by the Nereus consortium[16]) has a service that translates search statements in Spanish, French and German into English. This service uses in the background Google Translator for this purpose **(requirement 11)**.
  - VuFind seems to have very limited export functions to literature management software packages such as EndNote or RefWorks (**requirement 14)**.

[16] http://www.nereus4economics.info/

7.6.2 Manpower in the do-it-yourself scenario

The interviewees of the studied examples were asked to estimate the manpower that was used for the development and maintenance of their discovery services. The interviewees gave some indications and estimates of manpower (see table below). It is important to note that most interviewees mentioned that the manpower needed for the processing of metadata could vary enormously and was dependent on the quality and the description of the metadata formats used by the providers. Based on the indications in the table below, one could conclude that the development of the infrastructure for a national discovery tool by ABES itself will take 5 to 10 person-years and could be carried out in a development time of about one year. The bottleneck appears to be the acquisition and processing of the journal article metadata from the scholarly publishers worldwide: for this task no reliable figures for the manpower and cost involved are known.

	Manpower used (estimates)	**Development time**
Suchkiste	development VuFind: 4,5 person-years	Not available
	maintenance IT infrastructure: 4 hours/month	
Trove	development to first release: approx. 10 person-years	1 year, 4 months
	special effort (Trove stage 4) to improve coverage and access to Journal articles: 2 to 3 person-years	5 months
	maintenance: team of 2 developers, 1 user interface designer, 1 business analyst and a project manager (these team members are not full-time involved with Trove)	
University of Utrecht; University of Tilburg	maintenance of their discovery tools: up to 1 FTE/year	Not available
EZB link resolver	maintenance and development technical infrastructure: 1 to 1.5 FTE (an estimated 0.5 FTE for the link resolver)	Not available
Journals Online & Print	Not available	Less than 1 person-year
FINNA	IT development staff approx. 5 fte; another 12 FTE staff is involved in communication and training efforts with regard to the over 400 organizations that FINNA will serve in the future; rough estimate of the costs of the investment for hardware: 100.000 to 200.000 euro	10 to 15 person-years for the development of the present public interface (which will be further developed).

Table 2 Manpower estimates for a do-it-yourself scenario