The ABES Discovery study

A study into three scenarios for a National Webscale Discovery Tool for scholarly e-content

Maurits van der Graaf
Main report; 29-3-2013
Pleiade Management and Consultancy BV
Keizersgracht 62
1015 CS Amsterdam
The Netherlands
T: +31 20 488 9397
m.vdgraaf@pleiade.nl
www.pleiade.nl

Content

1. Introduction

1.1 Vision by ABES

ABES has recently developed a vision for a national resource discovery model for higher education institutes in France. An overview of decision is presented in the figure below.

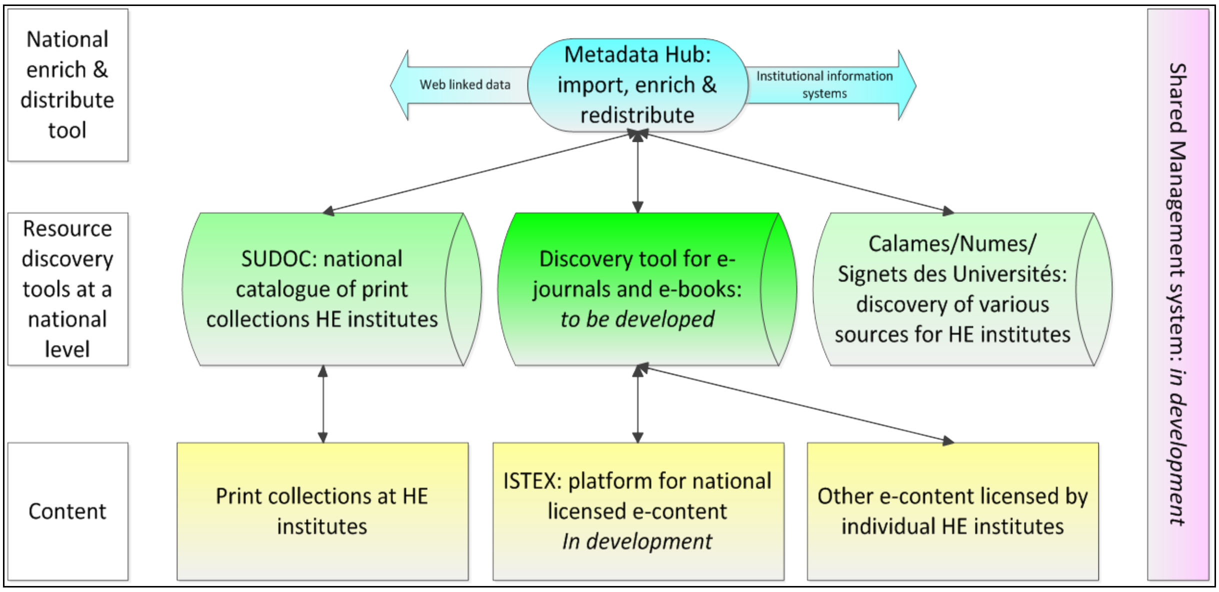

National enrich & distribute tool
Resource discovery tools at a national level
Content
Metadata Hub: import, enrich & redistribute
Web linked data
Institutional information systems
SUDOC: national catalogue of print collections HE institutes
Discovery tool for e-journals and e-books: to be developed
Calames/Numes/ Signets des Universités: discovery of various sources for HE institutes
Print collections at HE institutes
ISTEX: platform for national licensed e-content In development
Other e-content licensed by individual HE institutes
Shared Management system: in development

Figure 1 ABES vision

This vision builds on the well-established national catalogue for printed information resources SUDOC and recently developed discovery tools for archival collections (Calames), digitised corpora (Numes) and relevant websites (Signets des Universités).

ABES aims to expand these national resource discovery tools with a discovery tool for e-journals and e-books that are:

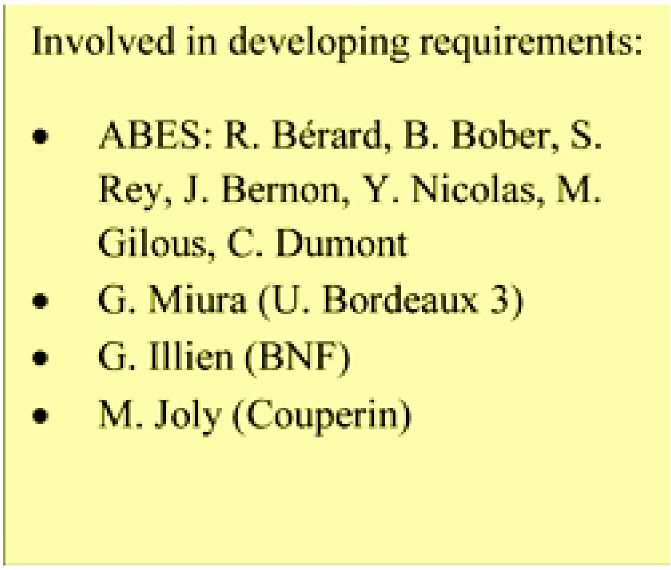
- Licensed and distributed at a national level (by national licences and distributed via the ISTEX platform)
- Licensed by individual higher education institutes.

The vision of ABES includes a Metadata Hub that will import, enrich and redistribute metadata to institutional information systems and to linked data in the World Wide Web.

1.2 Aims

The overall aim of the to-be-developed national resource discovery tool for scholarly e-content is to provide an exhaustive cartography of electronic documents in France. End-users will be able to access the data via a public interface and get to the retrieved documents in the most direct way possible (direct retrieval when permitted by copyright, otherwise delivery via interlibrary loan, pay-per-view and other mechanisms). In addition, it is foreseen that the national resource discovery infrastructure will enhance collaboration between French HE institutes in the development of collection management tools. With regard to the increasing collaboration between the libraries of the HE institutes, the first steps towards a shared management system have already been taken by ABES and its partners.

1.3 Scenarios and roadmap



Involved in developing requirements:

- ABES: R. Bérard, B. Bober, S. Rey, J. Bernon, Y. Nicolas, M. Gilous, C. Dumont
- G. Miura (U. Bordeaux 3)
- G. Illien (BNF)
- M. Joly (Couperin)

This report describes the results of a study into three possible scenarios to develop the above-mentioned national resource discovery service for e-content. This study was carried out by Pleiade Management and Consultancy with the following set up:

- **Phase 1 - Development of the requirements for the national e-content resource discovery service:** In this phase of the study, the requirements for the national e-content resource discovery service were developed in a workshop with key staff members within ABES, a number of interviews with other key players in the library world of France and a review and analysis of the recent literature about resource discovery for e-content.
- **Phase 2 - Exploration of existing or developing solutions at a national level in other countries:** in this phase, various existing or developing solutions at a national level in other countries where export by interviews and studying available documentation, websites and reports (see Appendix A and B).
- **Phase 3 - Exploration of the three scenarios:** the following scenarios were originally proposed by ABES and investigated in this study:
  - ***A do-it-yourself scenario, building a newly developed discovery service:*** This scenario is based on Trove, the online search engine developed by the National Library of Australia. In this scenario, the national discovery tool for France will be newly developed using open source software: the metadata and an index of the full text will be retrieved from each publisher. Users that are a member of the library with a subscription to the resource will be given immediate access. For others, a delivery mechanism should be provided (see Appendix A).
  - ***A scenario using discovery tools available on the market:*** A national resource discovery tool for all e-content resources covering all metadata, giving access to the full text versions of the scholarly articles or e-books via a link resolver mechanism. End-users can

use the discovery tool via a public interface. The national resource discovery tool in this scenario will be an adaptation of an existing discovery tool, such as Summon (Serial Solutions), Primo (Ex Libris), EBSCO Discovery (EBSCO) or OCLC WorldCat local (see Appendix B).

- ***A scenario that develops collaboration with Google Scholar:*** In this scenario, Google Scholar plays a central role, providing the public interface and the search engine. Access to full text is provided using a link resolver mechanism (see Appendix B).

The ultimate aim of this study is to set up a roadmap towards a national discovery tool for France, consisting of one of the above-mentioned scenarios or a combination of elements of these scenarios.

1.4 Set-up of the report

The report consists of four parts:

- The main text, describing the scope of the national discovery tool for France, the main results and conclusions of the three scenarios studies and the outline of the roadmap for further development.
- Appendix A, describing in detail the results of the do-it-yourself scenario study
- Appendix B, describing in detail the results of the scenario study with regard to the web scale discovery tools by libraries sister providers and this scenario with regard to a possible collaboration with Google Scholar.
- Appendix C, a table with a comparison of the interface requirements: VuFind, Primo, EBSCO Discovery, Summon, WorldCat Local and Google Scholar.

2. Requirements for a national discovery service in France

2.1 Introduction

In this chapter, first definition of a web scale discovery tool is given, followed by an overview of the landscape of discovery tools. Then, the scope of the national discovery tool for France is described. This scoping leads to the definition of 19 requirements.

2.2 Definition of a webscale discovery tool

Since a few years, webscale discovery tools are on the market for libraries. In addition, some libraries have developed webscale discovery tools themselves. Webscale discovery services can be described as a next stage in the development of library discovery services for end-users: earlier stages were OPAC catalogues and federated search engines. Webscale discovery services have the following characteristics[1]:

- Content:
  - The basis of a webscale discovery service is a vastly comprehensive centralised index (to the article level).
  - The centralised index is based on a normalised schema across content types
  - This index is created by harvesting content from:
    - local library resources
    - publishers and aggregators, that allow access to their metadata and/or full text content for indexing purposes
- Discovery:
  - Single search box providing a Google like search experience. (Frequently, there is also an advanced search interface)
  - Quick search results ranked by relevancy ranking that can be influenced by the library (for example giving local collections higher relevancy)
  - Options to refine the search (such as faceted navigation)
  - Filter on licensed materials from that particular library possible
- Access/delivery:
  - Includes digital resources at article level *and* the print resources of a library
  - Works with the library link resolver to provide access to the article level or document delivery service
  - Integration with functionality of the library catalogue so that the end-user can see if the retrieved print collection item is available and reserve it.
- Connected with the knowledgebase of the library/libraries involved: this characteristic applies only when the discovery services offers materials that are licensed by the library/libraries involved.

[1] Taken from Vaughan, 2012 with adaptions and additions

2.3 The landscape of webscale discovery tools

Four providers offer webscale discovery services as defined above: a combination of discovery portal (the interface) and a central index (the content):

- EBSCO's Discovery Service (www.ebscohost.com/discovery)
- Ex Libris's Primo Central Index (www.exlibrisgroup.com/category/PrimoCentral)
- Serials Solutions' Summon (www.serialssolutions.com/discovery/summon)
- OCLC's WorldCat Local (www.oclc.org/worldcatlocal)

The providers typically license the discovery portal and the central index as a unified package. However, a variety of discovery portals can be used to search the central indexes from EBSCO, Ex Libris, and Serials Solutions. There are several implementations that use VuFind or other discovery interfaces in conjunction with the vendors' central indexes[2].

Outside the commercial arena, there are several projects that have or strive to setup a webscale discovery service (mostly using open source software). One might categorize the various projects as follows:

- Focused on cultural heritage materials and/or open access materials:
  - **Europeana:** a large project for aggregating European cultural heritage content[3].
  - **Digital Public Library of America**: a project that will make the cultural and scientific heritage from the United States available, focusing on heritage and open access materials[4]. As both projects do not cover current scholarly literature, they are not further discussed in this report.
- Focused on current scholarly literature and/or cultural heritage materials and/or open access materials:
  - **Trove - the discovery tool in Australia**: this discovery service exists already for several years and gives access to current scientific literature as well as Australian heritage and cultural materials.
  - **Suchkiste, Journals Online & Print, and EZB**: Suchkiste is a German discovery service for scholarly content that is nationally licensed. In addition, the Journals Online & Print service – based on the data from the Electronic Journals Library (EZB) and the ZDB – the union catalogue for journals in Germany - is also a relevant service indicating availability of the full text for end-users. In addition, the EZB offers the EZB linking service that is also relevant to this study.
  - **The National Digital Library in Finland (FINNA)**: a project to create a joint public interface for materials and services of libraries, archives and museums. This project includes current scientific literature and cultural heritage materials.

[2] The ins and outs of evaluating webscale discovery services, Athena Hoeppner, Computers in Libraries; Vol 32, No 3 - April 2012

[3] Europeana, Business plan 2012

[4] The Digital Public Library of America; concept note; March, 2012

2.4 The positioning of a national webscale discovery tool in France

The envisaged national discovery tool should serve three categories of Higher Education institutes and their libraries with regard to discovery *and* access for their end-users to digital *and* print collections:

- **French HE libraries with installed discovery tools:** an estimated 20 to 25 French HE libraries have implemented webscale discovery services from the various providers. The envisaged national discovery tool should in principle offer such functionality that these HE libraries might exchange (in the longer-term) their presently-used discovery tool for the national discovery tool. However, the national discovery tool should also be able to interact with discovery tools of the above-mentioned libraries in such a way that it immediately will give advantages for the users from these institutes (for instance with regard to coverage of French scholarly content). This specific requirement – sharing (parts of) the index or metadata with other discovery services – will be referred to as **requirement 1**.
- **French HE libraries with link resolvers/knowledgebases implemented:** many other larger HE institutes and libraries in France have implemented link resolvers and knowledgebases from the various providers. The national discovery tool should be able to interact with those link resolvers and knowledge bases in order to provide the end-users of these HE institutes with proper access to the digital collections of their libraries. This positioning requirement leads to **requirement 2** that will be described in detail in paragraph 2.4.
- **French HE libraries without link resolvers/knowledgebases implemented**: the national discovery tool should offer a discovery service for libraries without a local link resolver/knowledgebase (most often smaller HE institutes that might never implement link resolvers). In other words, the discovery tool should provide some sort of access service to end-users of these libraries. This positioning requirement leads to **requirement 3** that will be described in detail in paragraph 2.4.

2.5 Interoperability requirements

The national discovery tool is considered an important building block in the present and future national library infrastructure in France. Therefore, the national discovery tool should be interoperable with the following existing library systems:

- **Knowledgebase/link resolvers already installed at French HE libraries**: interoperability with link resolvers is important to give access to the holdings of the library, with which the end-user is affiliated for two reasons: (1) to provide a link to the appropriate copy of that library and (2) to provide a filter on the library holdings so that the end-user can limit the search results to the holdings. This requirement will be referred to as **requirement 2** (see also paragraph 2.3)**.**
- **Integration with the national union catalogue SUDOC**: the national discovery tool needs to be integrated with the SUDOC catalogue in order to point the end-user to the holdings of his/her own library and/or other libraries. In the text box, the presently available mechanisms to achieve this are presented. This requirement will be referred to as **requirement 3** (see also paragraph 2.3).
- **Integration with local OPACs and ILL:** the national discovery tool needs to be integrated with the local OPACs of individual HE libraries in order to provide the end-user access to the circulation of print items (availability, reserve the item etc.) and/or the interlibrary loan service (ILL) in order to enable the end-user to order the document directly from the discovery service environment. This requirement will be referred to as **requirement 4**.

Sudoc interoperability mechanisms
Information about possible mechanisms to integrate data from the French union catalogue Sudoc:

- Sitemap crawling, see http://www.abes.fr/Acces-direct-a/Pour-les-developpeurs (Sudoc)
- Documentation: http://www.sudoc.fr/noticesbiblio/sitemap.txt and http://www.sitemaps.org/
- RDF crawl (the URLs are in the sitemaps)
- SRU
- Exports made internally (requiring a lot of work for ABES)

Sudoc could be updated from third-parties application with:

- SRU update
- OAI-PMH
- Manual imports

In addition, the national discovery tool should also be interoperable with a number of national library systems that presently are in development:

- **The knowledgebase of the future shared library management system (with ERM functionality) in the cloud**: ABES is presently working with a number of HE libraries in France on developing requirements for a shared library management system. It is envisaged that this shared library management system will be ‘in the cloud’ and consists also of electronic resource management functionality with a knowledgebase. It is envisaged that this knowledgebase will use the standards ONIX-PL and KBART that are presently being developed. The development of this national shared library system is planned for 2013-2014. This requirement is referred to as **requirement 5**.
- **ISTEX platform with nationally licensed content:** The ISTEX project (Initiative d ’excellence de l ’Information Scientifique et Technique – see www.istex.fr) is carried out by four parties: CNRS, ABES, Couperin and the CPU. The Agence Nationale pour la Recherche (ANR) has financed a project with 60 million Euros for three years: 55 million Euros for the acquisition of content and 5

million Euros for the development of a platform to store the content. ISTEX aims for a centralized archive with scholarly information that is acquired by national licenses in order to offer a retrospective collection for the scholarly community in France. The first national licenses[5] are planned to be signed in the course of 2013. At first, access will be given via the sites of the publishers. The platform will be developed from May 2013 to May 2014. From then on, access to content will be given via the platform. The development of the national discovery tool will be closely aligned with the development of ISTEX platform. This requirement will be referred to as **requirement 6**.

- **The planned Metadata Hub in order to have metadata enriched and redistributed:** ABES presently is working on the development of a prototype for a Metadata Hub. The idea is that this would function as a metadata factory to get good metadata out of diverse and messy data, to deduplicate metadata and to redistribute the cleaned and enriched metadata to relevant systems. The national discovery tool will be such a relevant system. One of the goals is not to be totally dependent on (international) knowledgebase providers, especially with an eye on French content. ABES strives to make metadata as open as possible. Therefore, ABES has recently decided to apply a French license[6] that is compatible with the Open Data Commons Attribution License (ODC-BY). This means that that all French HE libraries that will produce metadata will do so under this license and therefore those metadata can be included in the discovery tool. This requirement will be referred to as **requirement 7**.

[5] Two national licenses are already realized in an earlier phase, see www.licensesnationales.fr . For these licenses, the IP address ranges of 125 HE institutes have already been collected by ABES for authentication purposes.

[6] ABES will use the open license/license ouverte Etalab (http://www.etalab.gouv.fr/pages/licence-ouverte-open-licence-5899923.html)

2.6 Index and coverage

The national discovery service by ABES focuses on scholarly content: in the first place e-journals, in the second place e-books (title and chapter level), other scholarly content and in the longer-term access to other media types (for example enriched publications). Of special importance is access to French scholarly e-content: the national discovery tool – together with the Metadata Hub that is in development – will form an instrument to enhance access to French language e-content and to e-content by French publishers. This has led to the following requirements, summed up in table 1 below. The following remarks with regard to these requirements are important:

- **Emphasis on e-journals:** It is important to note that with regard to coverage of the discovery tool, the emphasis lies on scholarly e-journals and e-books as a second priority. Other types of scholarly content are seen as important, but to a lesser extent. It is conceivable that some libraries would like to add local digital content: some to make it publicly available, some to make it only available to their own campus[7]. Functionality with regard to this is desirable.
- **Emphasis on metadata:** with regard to what should be indexed, the metadata are seen as the most important part, as are citation links. Full text indexing is seen as desirable. Inclusion of user contributed reviews has a lower priority.

Index and coverage requirements	
Levels of metadata indexed (requirement 8):	
Basic metadata	Yes
Keywords, subject descriptors	Yes
Abstracts	Yes, preferably
Full text ('deep index')	Yes, preferably
Citation links	Yes, preferably
User contributed reviews; reviews from vendors such as Amazon	Considered not necessarily
Coverage (requirement 9):	
Scope of the coverage	The emphasis lies on e-journals and their articles
Publishers	Emphasis on e-journals and their articles
SUDOC and/or local catalogue with print collection	Yes, see paragraph 2.4
E-book vendors	Yes, preferably title and chapter level indexing
OA journals, repositories (HAL), Google books/Hathi Trust etc.	Yes, emphasis on e-journals and e-journal articles
Local content; 'private' digital content defined by participating libraries	Desirable, see text

Table 1 Requirements with regard to index and coverage

[7] This can be especially relevant for the full text of theses for some universities.

2.7 Interface

Search (requirement 10)	
Single search box	Yes
Advanced search option	Yes
Languages of the interfaces	French, English, and one other language: Spanish, German or Italian
Non-English language support (**requirement 11**): Spelling suggestions Search term translator from non-English language to English Sorting option on language	See text below
Mobile interface	Yes
Local library customisation/ local branding	Yes
Recommender options (requirement 12)	
Recommender functions	Yes, preferably (see text below)
Presentation of the results (requirement 13)	
Relevancy ranking	Yes
Faceted navigating/search refinement options	Yes
Export options (requirement 14)	
EndNote	Yes
Plain text	Yes
Structured format	Yes
Csv	Yes
Sorting options (requirement 15)	
Publication date	Yes
Author	Yes
Source title	Yes
Relevance	Yes
Number of citations	Yes
Language	See below
User accounts (requirement 16)	
Save results	Yes
Save searches	Yes
Alert services	Yes
Social features (requirement 17)	
Sharing results	Yes
User tagging	Yes

Table 2 Requirements with regard to the interface

In the table above, the main requirements are presented for the interface. All requirements are within the normal range of modern interfaces for search engines. However, there is one additional wish that could give added value to the end-users in the French HE institutes. Spelling suggestions in discovery are presented for the English language, and not for the French language. It is conceivable for a number of end-users there is a language barrier with regard to searching and reading English language. Therefore, the following options are desirable:

- **A search term translator** that will translate French language search terms into English language search terms
- **A sorting option** on the results page, that enables the users sought on the language of the article.
- **Spelling suggestion** feature that also works well in the French language.

Other remarks:

- **Social features (Web 2.0)** are deemed less important: scientists who want to use such features will use dedicated platforms such as Mendeley.
- **Faceted navigation/search refinement** options are crucial for end-user searches and are dependent on good metadata, which presents another possible role for the Metadata Hub.
- **Recommender functionality** is very much in development and is seen as of increasing importance for end-users. For the national discovery tool, an article recommender service is desirable: 'users that retrieved this article also retrieved these articles'.

2.8 Other functionality

The national discovery tool should also have the following requirements:

- **Open API platform:** It is desirable that the platform of the national discovery tool will have an application programming interface (API). An important function of such an API (or other mechanism) is to open-up French e-content for other discovery services and/or Internet search engines. This requirement will be referred to as **requirement 18**.
- **User statistics:** It is important that the national discovery tool will provide statistics on the usage of the discovery tool, as well as nationwide as per participating HE library. This requirement will be referred to as **requirement 19**.

2.9 Overview requirements

In the table below, an overview of the requirements with regard to national discovery tool - as discussed in the paragraphs 2.3 to 2.7 is presented.

	Overview requirements	**Purpose:**
1	Sharing (parts of) the index or metadata with other discovery services	Interoperability with already installed discovery services and opening-up (French) metadata for internet search engines
2	Interoperability with local link resolvers/knowledge bases	Access to e-resources
3	Interoperability with union catalogue in order to give availability information	Supporting access to resources for libraries without linkresolver via SUDOC
4	Integration/interoperability with local OPACs and ILL service	Supporting access to circulation functionality with regard to p-resources
5	Interoperability with the knowledgebase of the future shared library management system	Future-proof building block of national library infrastructure in France
6	Interoperability with a platform with nationally licensed content	Future-proof building block of national library infrastructure in France
7	Options to deduplicate, enrich and redistribute metadata	Relation to Metadata Hub
8	Specifications with regard to the metadata and/or full text indexed	Scope of discovery service
9	Specifications with regard to the coverage of the scholarly content and the option to add 'private' content to the index	Scope of discovery service
10	Search options	User experience
11	Non-English language support	User experience
12	Recommender options	User experience
13	Presentation of the results	User experience
14	Export options	User experience
15	Sorting options	User experience
16	User accounts	User experience
17	Social features	User experience
18	Open API platform	Interoperability with regard to opening-up (French) metadata for internet search engines
19	User statistics	Management of the discovery service

Table 3 Overview requirements national discovery tool

These 19 requirements for a national discovery tool for France are used in the following ways:

- To judge the outcomes of the do-it-yourself scenario study (see appendix A)
- To judge the outcomes of the scenario study with regard to the discovery tools by the library system providers and the outcomes of the scenario study with regard to Google scholar (see appendix B)
- To compare the interfaces/portals of the various discovery systems, see appendix C.

3. Exploration of the three scenarios

3.1 Components of a national discovery system

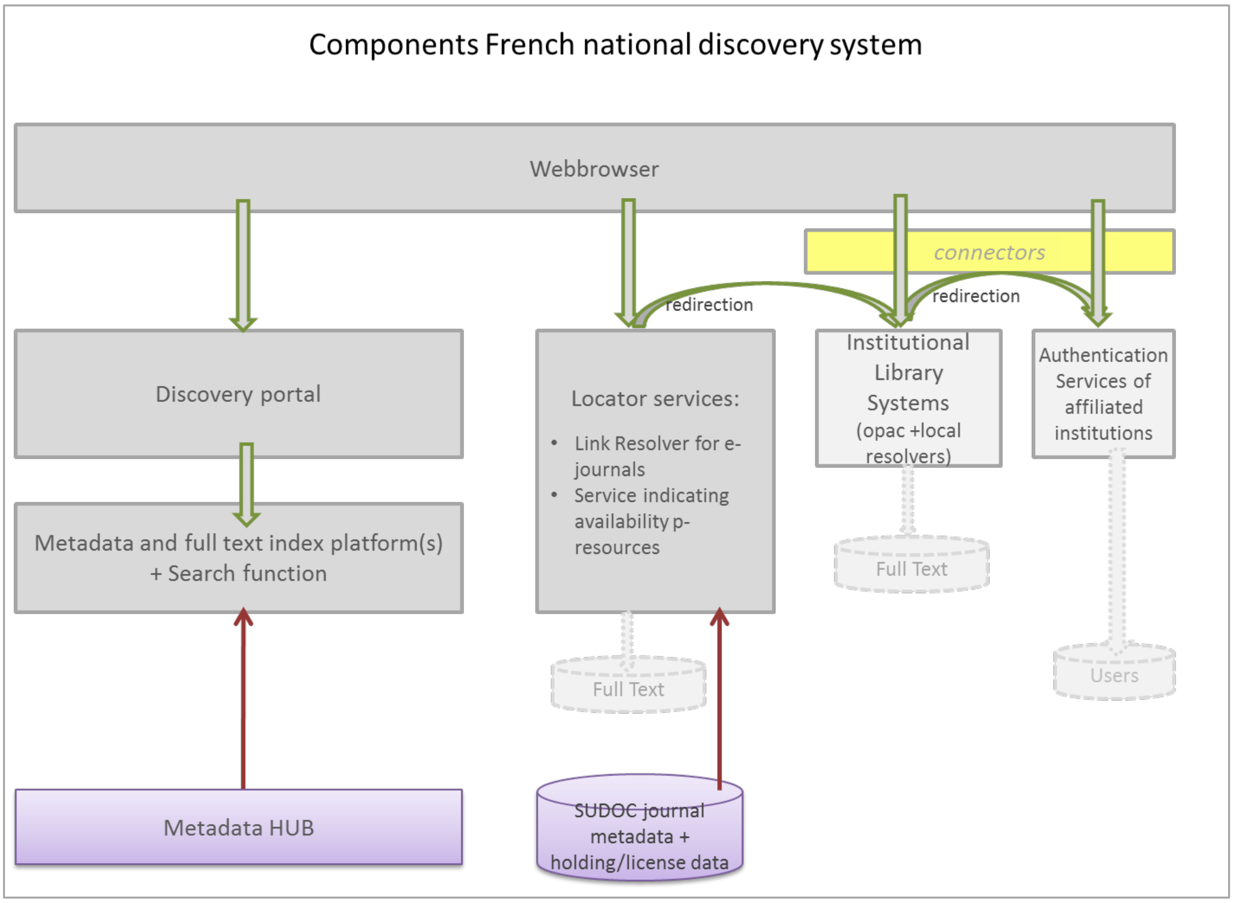

Components French national discovery system
Webbrowser
connectors
redirection
redirection
Discovery portal
Metadata and full text index platform(s) + Search function
Locator services:
Link Resolver for e-journals
Service indicating availability p-resources
Institutional Library Systems (opac +local resolvers)
Authentication Services of affiliated institutions
Full Text
Full Text
Users
Metadata HUB
SUDOC journal metadata + holding/license data

Figure 2 Components discovery system

From the exploration of the scenarios, the following main components of a national discovery tool can be identified (see figure 2):

- **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.
- **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).

3.2 Results of the do-it-yourself scenario study

3.2.1 Introduction to the do-it-yourself scenario

In this scenario, the national discovery tool for France will 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: these are described and discussed in Appendix A. In this chapter, the main outcomes of this scenario study will be presented and discussed. Especially the case reports of Trove – a discovery service built by the National Library of Australia, of FINNA - a discovery service in development by the National Digital Library of Finland, the discovery tool Suchkiste and the Journals Online & Print webservice and the EZB linking service from Germany are very relevant for this scenario.

3.2.2 Main results of the do-it-yourself scenario study

- **Portal:**
    - With regard to the portal, 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. The results are listed in Appendix C. Clearly, the most important requirements are met by VuFind with two exceptions:
        - Presently, VuFind has very limited export functions to literature management software packages such as EndNote or RefWorks (requirement 14).
        - The specific non-English language support functions that are considered desirable to support the French language were not observed (requirement 11). FINNA has 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[8].
- **(Meta)data platform:**
    - The self-built discovery services that produced their own centralised indexes for scholarly literature only have achieved a very limited coverage. The experiences in this respect of Trove and Suchkiste make it clear that building a centralised index of a *selection* of the worldwide scholarly literature is feasible. However, to build a centralised index of the

[8] Economists Online (a portal for economics literature by the Nereus consortium 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.

    entire world wide scholarly literature is seen as a major effort, requiring a lot of manpower. The respondents 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. This is one of the reasons why FINNA included next to its own index with metadata of Finnish materials the index of Primo Central Index from ExLibris for its coverage of the scholarly literature. FINNA combines these two indexes with their portal software VuFind and provides a seamless integration of those two indexes for the end-users. Suchkiste provides another example of a combination of two index platforms as the Suchkiste metadata platform for national licences in Germany can be seamlessly combined with the Primo Central index for Primo clients.
    - Full text indexing is not reported by the self-built discovery tools studied here as well as the inclusion of citation links.
- **Locator services:**
    - The EZB link resolver - based on the data of the EZB union catalogue for e-journals - provides a very good example of a national link resolver. The EZB link resolver can be used by the libraries without link resolver of their own to support the discovery by their end-users. In addition, the EZB link resolver can be used by libraries with a link resolver of their own: either by using the EZB link resolver as a target for their own link resolver, or by using the knowledgebase data of the EZB for their local knowledgebase that supports their local link resolver.
    - The Journal Online & Print webservice (JOP) provides an (probably unique) service for end-users with regard to the print journals holdings of their library. Based on the German union catalogues ZDB and EZB, this webservice can indicate in search engines the availability of a journal article in the print holdings of the library of the end-user. In combination with the EZB link resolver these two services support the end-users in the location of the full text in their library.
- **Connectors:** FINNA puts great effort in integrating the functions of the local library catalogues into FINNA, such as reserving or borrowing an item in the local library collection. The ultimate aim of FINNA is to replace all local front-ends with FINNA. The VuFind portal software (possibly after additional development efforts) is capable of supporting these services. However, for the French national discovery service a replacement of the local OPACs is not envisaged.
- **Resources needed:** in appendix A (paragraph 7.6.2) an overview is given of estimates by the respondents of the manpower needed for their self-built services. The discovery services Suchkiste and Trove mention 5 to 10 person-years for the development of their discovery services. With regard to maintenance, Trove has a team of 5 individuals in place that also has other tasks. Suchkiste only mentions the technical maintenance (6 person days per year) but gives no estimate for processing efforts for new national licences. FINNA mentions 10 to 15 person-years for development up until the present stage of development. The locator services EZB link resolver and the Journal Online & Print webservice mention both 1 person-year and less for development and maintenance.

3.3 Results of the scenario studies with existing discovery services

3.3.1 Introduction to the two scenarios with regard to existing discovery services

In this chapter, the main results of the study into two scenarios are presented. The full results are described and discussed in Appendix B.

One scenario study investigated the options in using the four discovery tools available on the market for developing the French discovery tool. These discovery tools are produced by for library system providers: Summon (Serial Solutions), Primo (Ex Libris), EBSCO Discovery (EBSCO) or OCLC WorldCat local. The national resource discovery tool in this scenario would be an adaptation of an existing discovery tool. Another scenario study looked into the possibilities to set up a French national discovery tool in collaboration with Google Scholar. In this scenario, Google Scholar would play a central role, providing the public interface and the search engine. Access to full text is provided using a link resolver mechanism.

3.3.2 Main results

- **Portal:**
    - All discovery services provide a limit option on the language (via facet mostly).
    - Google Scholar, Primo and Summon have spelling suggestions for the French language available. Summon has automatic pluralisation and treatment of compound words in French language (and other languages). In addition, Primo offers multilingual thesaurus support.
    - None provide search term translation functionality. Google Scholar indicated that this was possible to build (using Google Translator); however there was doubt if this would fulfil an important need in the French end-user community.
    - With regard to the other requirements, the interfaces are more or less comparable (see Appendix C for a full comparison).
- **(Meta-)data platform:**
    - **Coverage:** All providers claim to cover the worldwide scholarly literature extensively and tool index the full text of it for a large part of it. Google Scholar even states that they cover the full text for the large majority and only in exceptional cases solely the metadata. In this respect, Google Scholar also seems to have a more strict definition of scholarly literature than the other providers. This comes especially to the fore when Google Scholar indexes union catalogues and/or link to Google Books: this is only done for scholarly books and not for other publications that are often included in collections of academic libraries (e.g. novels, newspapers). With regard to the coverage of scholarly journal literature, there is reason to believe that in the longer term existing differences in coverage between the various discovery tools will vanish as publishers will increasingly distribute their data to other discovery systems as well. With regard to the metadata quality, all discovery systems have mechanisms in place to use metadata from A&I databases to enrich the metadata delivered by the primary publishers via match & merge mechanisms.

  - **Other platforms:** Ex Libris is the sole provider with a policy to connect their centralised index to other Solr platforms with the so-called deep search connection. All other providers prefer to receive the data to include it in their centralised index.
  - **Enrichment:** Enrichment and redistribution to other parties of metadata from a centralised index by Metadata Hub is generally prohibited by the license conditions by the primary publishers. Enriched metadata from an open platform would be included in the centralised indexes of the discovery tools via match & merge mechanisms.
- **Locator services:**
  - **Interoperability:** Interoperability between link resolvers of other providers and the studied discovery services pose no problem. Google Scholar can integrate all link resolvers as well.
  - **Knowledgebase:**
    - All library system providers participate in the development of the KBART standards for knowledgebase data. Therefore, it can be expected that the exchange of knowledgebase data will be facilitated in the near future by using this standard. However, the present experiences show that an important percentage of sources will not match (10 to 30%, see also paragraph 4.4.2 of Appendix B).
    - Google Scholar is creating a sort of knowledge base of its own by asking publishers to provide the holding data of the libraries that are their customers. Google Scholar approaches the consortia to ask if they will include such a delivery from the publisher to Google Scholar in their licence conditions.
  - **Location of print sources:**
    - All providers – including Google Scholar – have many examples of integrating/connecting to union catalogues[9].
- **Connectors:**
  - One of the strongest points of the discovery services by the library system providers appears to be the connectors to the OPACs of many different library management systems. These connectors bring the end-user to functionality such as reserving of print items, often within the environment of the discovery tool. This could enable libraries to replace their OPAC with the discovery service. Each connector to the OPAC has to be established separately, the effort to create these connectors for over 150 institutional systems however is considered to be rather limited (in terms of perfume person months).
  - Google Scholar connects to OPACs of individual institutes via WorldCat or Sudoc.
- **Resources needed:**
  - **Manpower:** The case report of RERO - the Swiss network of libraries (see appendix B, chapter 2) - demonstrates that using a commercial discovery service (in this case Primo)

[9] The German webservice JOP (see Appendix A 5.3 for a full description) that indicates the availability of a journal article in the printed version of a journal appears to be unique.

for group of libraries also requires manpower from the side of the libraries. RERO reported for their situation a project team of 7 members that were working on the development for over 6 months, although not all team members were working full-time project. For maintenance, one estimates 0.5 to 1.0 FTE for the administrative efforts on behalf of the library network. It has to be noted that RERO aims in the longer-term to replace its OPACs by the discovery service.

- **Costs:** The possible costs of a national discovery service provided by the one of the library system providers based on an adaption of their discovery tools was not discussed with the respondents because of legal reasons. The financial costs can be divided in development costs and maintenance costs. Based on what is known from existing subscriptions to webscale discovery services by individual libraries, it is clear that the maintenance costs for a national discovery service by one of the commercial providers would be in the order of at least several hundred thousand euros per year. This would mean that ABES would have to develop a businessmodel that probably will have to include contributions from the participating HE libraries. The feasibility of such a business model is considered doubtful. In case of collaboration with Google Scholar, the maintenance costs can be expected to be much lower and possibly fit in within the budget of ABES.

3.4 Conclusions from the scenario studies

3.4.1. Metadata Hub as a component in webscale discovery

An important function of ABES is to fulfil a national role as an expert on metadata. It is foreseen to further built on this role by an operational component named the Metadata Hub. A similar function is for instance seen in the FINNA set-up (with the name Record Manager; see appendix A). However, from the studied examples, the acquisition and processing of metadata of the international scholarly journal literature appears to be a major bottleneck of a do-it-yourself scenario. Because of the long tail of publishers (see table 4), the efforts involved in acquiring journal article metadata and processing them are making it very labour-intensive to achieve an adequate coverage of the scholarly journal literature and are therefore deemed not to be feasible for the French national discovery tool.

Publisher size	Number of journals published	Scopus figures 2009	Scopus figures 2009[10]
		% journal articles published	% publishers
very small	1 - 10 journals	30.9%	97%
small	11 - 50 journals	14.6%	2%
medium	51 - 250 journals	6.9%	0.32%
large	250 - 1000 journals	6.2%	0.04%
very large	> 1000 journals	41.4%	0.08%
total	17565	1628354	4993

Table 4 The long tail of journal publishers

For this reason, it is envisaged to create a French metadata platform with selected content that is especially relevant for the French scholarly community. Such a French metadata platform could contain metadata of the national licences, content from French publishers and other relevant selections from the scholarly literature. The metadata for the French platform will then be enriched by the Metadata Hub. The metadata at the French platform will be Open Access and can be used by other services to enrich their own metadata via match & merge mechanisms. From the scenario studies it has become clear that all discovery services used these mechanisms to enrich their metadata. The Metadata Hub project (see also figure 3) is currently under way at ABES and will provide a proof-of-concept prototype mid-2013. The project already foresees a platform for distributing the enriched metadata in various formats.

The results of this study indicate that this metadata platform by the Metadata Hub can have a function in improving the discovery by the French HE community via existing discovery systems.

[10] Based on 2009 data from the Scopus database , table adapted from Eefke Smit, Maurits van der Graaf; Journal article mining: the scholarly publisher's perspective; Learned Publishing, 25: 35 to 46; doi:10.1087/20120106

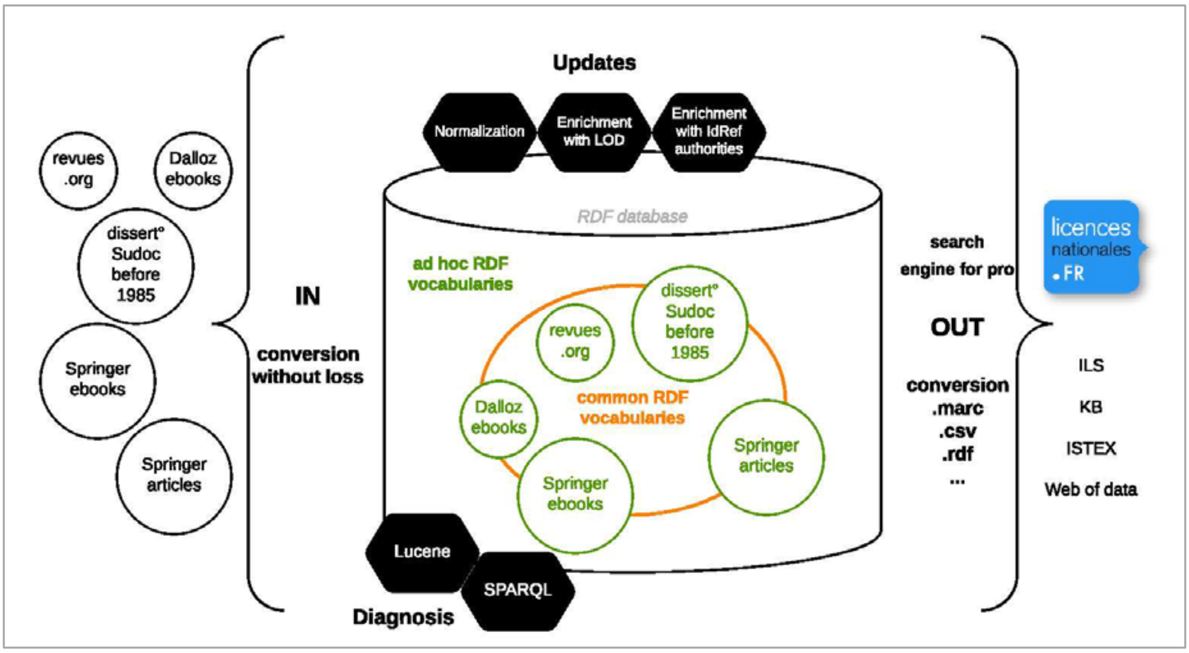


Figure 3 Metadata Hub by ABES

3.4.2 A national locator service in webscale discovery

A major outcome of the do-it-yourself scenario study was the possibilities to set up a national locator service along the lines of the EZB link resolver and the JOP web service. From the study of the existing discovery services by the library system providers it became clear that such a national discovery service also has to be newly developed in this scenario. Only Google Scholar appears to be working on a system that - if successful - might bypass the need for such a national locator service in the longer-term.

A national locator service can function as an important improvement in the discovery by end-users is in the French Higher Education community using existing discovery systems.

3.4.3 The other components discussed

With regard to the other components of the national discovery tool the following can be concluded:

- Portals: the existing portals - as compared in Appendix C - appear to be rather comparable and differ only with regard to the French language support functions. A new-to-be developed portal will have limited added value and will require a larger effort to change the present behaviour patterns of French HE users in order to attract sufficient usage.
- Connectors: Sudoc and WorldCat are already integrated in most webscale discovery tools and via this way can provide end-users with insight in the print collections of the libraries and in a number of instances also direct access to the OPAC of their libraries. A new-to-be-developed national discovery service would not have added value in this respect. Also, when the envisaged shared ILS in the cloud will be developed, these connectors will become obsolete.

3.4.3 Integration of the two to-be-developed components in existing discovery systems

Benefits for various target groups	**enriched metadata by Metadata Hub**	**national locator service (see explanation in text)**
HE libraries with discovery tool	Yes (1)	Yes (2)
HE libraries with local link resolver	Yes (1)	Yes (3)
HE libraries without local link resolver	Yes (1)	Yes (4)
HE end-users using free search engines such as Google Scholar	Yes (1)	Yes (5)

Table 5 Benefits of the 2 components

Based on the above-mentioned conclusions about the two components and with an eye on the limited resources of ABES, it is proposed to set-up a roadmap for the development for these two components. This implicates that ABES will not strive to build a new national discovery service but align/integrate those two components into existing discovery services.

In the table 5, the benefits for the various target groups are presented:

1) Webscale discovery services such as Primo, EBSCO Discovery, OCLC WorldCat Local and Summon but also Google Scholar will benefit from the enriched metadata by Metadata Hub.
2) HE libraries that subscribe to a webscale discovery tool will therefore benefit from the above-mentioned enriched metadata.
3) HE libraries that have a local link resolver can make use of the national locator service in two ways: they can use the national locator service as a target for their local link resolver or use the knowledgebase data from the national locator service for their own local link resolver.
4) HE libraries without the link resolver will obviously benefit from the national link resolver by integrating it in any database they subscribe to.
5) HE end-users that use free search engines such as Google Scholar, Microsoft Academic Search, Scirus and PubMed will benefit from the national locator service by getting access to electronic materials that are licensed by their libraries and/or by getting information about the availability of the print resources of their libraries.

4. A roadmap for improvements in discovery of scholarly literature in France

4.1 A roadmap for implementation of the discovery strategy

In this chapter, the outlines for a roadmap for improvements in the discovery of scholarly literature in France are presented in three (partly parallel) steps.

Step 1: Development of Metadata Hub

The Metadata Hub project has been started already, independently from this study. However, the outcome of this study means that it is foreseen that the metadata for selections of scholarly literature (such as the national licenses) enriched by the Metadata Hub would improve the discovery of this literature via webscale discovery systems. Therefore, it is recommended to proceed with high priority with this project to enrich metadata of selections of scholarly literature that are important to the French Higher Education community. Its main functions are:

- Aggregation of metadata of selections of the scholarly literature:
    - Examples of selections of the scholarly literature are the metadata of the content of the national licences, metadata of scholarly publications by French publishers.
    - Criteria for selection are:
        - important to the French HE community
        - neglected by other parties enriching metadata
        - useful for discovery systems, knowledgebase systems
- Analysis, cleaning and enrichment of those metadata, thus improving their quality and their discoverability. Examples of cleaning and enrichment are:
    - diagnose vital problems in the metadata and inform the primary publisher about these
    - generate out of the actual articles the journal holdings in order to identify gaps (relevant for knowledgebases and discovery systems
    - to add English-language and French-language descriptors for efficient faceted searching and browsing in discovery systems
    - to add international and/or national authority data
    - to add links
- Redistribution of the improved metadata by
    - conversion and making available for redistribution in various formats such as marc21, rdf, json by (1) data dumps for harvesting (2) web services and (3) linked data

The Metadata Hub will use RDF as the common data model. Metadata Hub works with Virtuoso, OpenRefine and SILK as automation tools.

Step 2: Development of a national locator service, including a national knowledgebase

As a second step, but in parallel with step 1, ABES will start a development of a national locator service which also implies a link resolver using national knowledgebase data. In this respect, the Knowledgebase + project of JISC and the GOKb initiative (see also Appendix A, paragraph 3.2) are very relevant:

- Knowledge Base+ is a recently developed 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 in first instance on data on JISC licensed content (NESLi2, SHEDL and WHEEL agreements) 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.

- The global open knowledgebase (GOKb) aims to become an open knowledgebase using standards-based architecture and with a CC0 license. The initiative is part of the Kuali Ole open source library management system development. The partners of Kuali Ole - over 20 American academic libraries - work together with JISC (Knowledgebase +) on this project. The aim is that the GOKb will interact with Knowledgebase + and other collectively managed knowledge bases.
  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.

GOKb data elements
title description
standard ID
package (a.k.a. collection)
platform
local ERM system data elements
subscription (deal)
purchase order
issue entitlement
license
usage statistics

The development of the national locator service means that a national knowledgebase along the lines of Knowledgebase + should be developed for the French HE libraries. In figure 4 the potential data flows from Sudoc, GOKb and a French national KB are depicted.

For this national locator service, the following items are (minimally) necessary:

- **Print holdings data**: these data are available in SUDOC
- **E-holdings data**: for the national licenses, these data are already available at ABES. Other data have to be collected from publishers via Couperin and from individual institutes in a collaborative effort with the French HE libraries, possibly aligned with the GOKb initiative and based on the Knowledgebase + experiences by JISC. Workflow and quality control procedures have to be developed so that all parties involved benefit from the knowledgebase data.
- **IP address ranges per library**: these data are already available at ABES in relation to the national licenses
- **Library identifiers:** codes to identify libraries are already available in SUDOC

- **Link resolver software and webservice software**: according to the experiences of EZB and JOP, the development of the software for the link resolver and webservice is a relatively minor effort (less than 1 person-year).

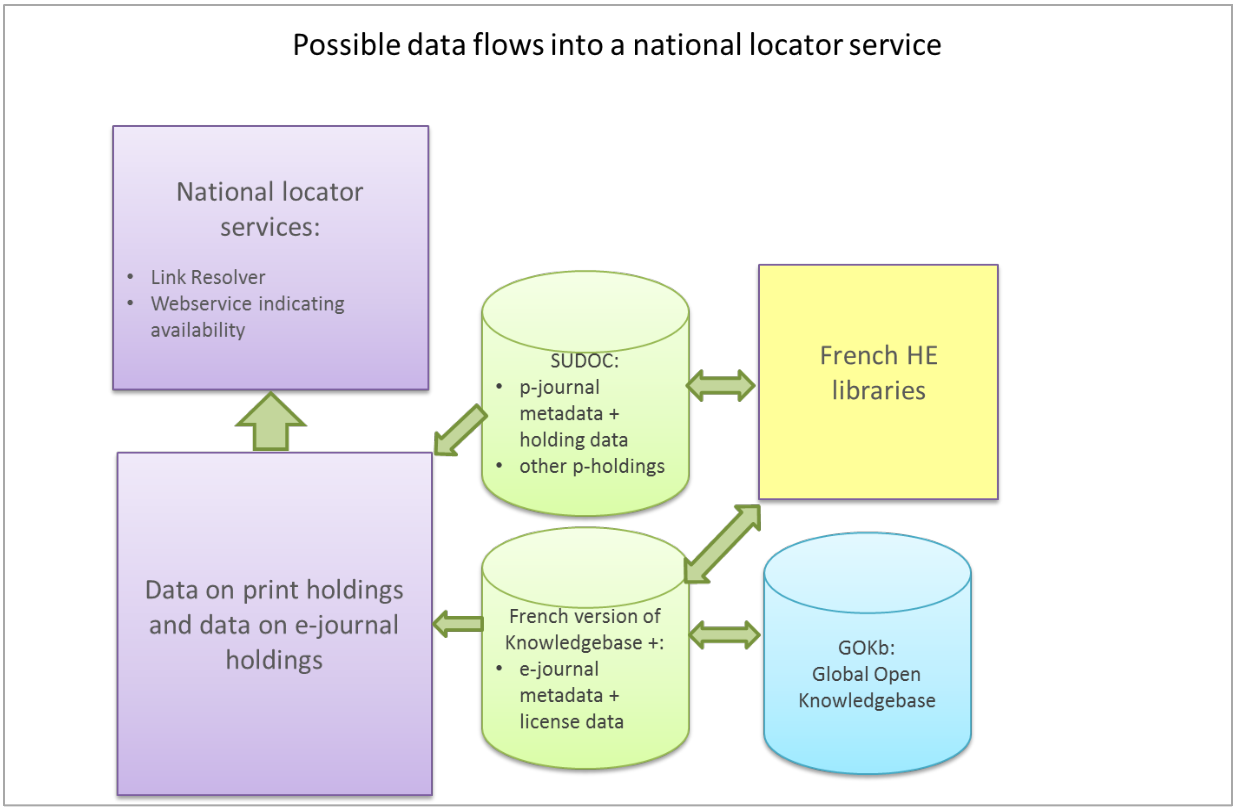

Possible data flows into a national locator service
National locator services:
• Link Resolver
• Webservice indicating availability
Data on print holdings and data on e-journal holdings
SUDOC:
• p-journal metadata + holding data
• other p-holdings
French version of Knowledgebase +:
• e-journal metadata + license data
French HE libraries
GOKb: Global Open Knowledgebase

Figure 4 Data flows into a national locator service

For the development of the national locator service and the national knowledgebase data collection the following phases are envisaged:

- In the first half of 2013 ABES will explore in close cooperation with Couperin and ADBU the options to collaborate with GOKb and Knowledgebase + in order to set up a national knowledgebase data collection for France.
- ABES will develop a prototype for a national locator service based on SUDOC data for print holdings and on GOKb data for some digital holdings in October 2013. For this prototype one will develop a web service along the lines of the German Journals Online & Print webservice integrated with an Open Source link resolver that will use (a probably limited set of) GOKb data. It is important to note that the Metadata Hub will be used to validate these GOKb data.
- Subsequent development of the national knowledgebase will focus on:
    - Collaborative working procedures at a national level with Couperin for the consortia licences, with ABES for the national licences and the French HE libraries for the individual licences. These collaborative working procedures will ensure that the knowledgebase data not only will feed into the national locator service but also will facilitate the efforts by the French HE libraries to maintain/update their knowledge bases as part of their ERM systems. It is also important to note that this effort with regard to e-holdings can be seen

    as an extension of the collaborative activities with regard to the union catalogue SUDOC, now primarily focused on print holdings.
    - Collaborative working procedures at an international level with GOKb partners.

- Subsequent development of the national locator service will focus on:
    - Further refinement and improvement of the locator services
    - Integration of the national locator services with free discovery tools, integration with the local link resolvers of HE libraries as a target and integration with subscribed search engines & databases of HE libraries without a local link resolver.

Step3: Active approach by ABES to integrate the enriched metadata and the national locator service in existing discovery services

An active approach by ABES is envisaged in the implementation of both the enriched metadata by the Metadata Hub and the national locator service. This will mean among others:

(1) An active approach of existing discovery services with the eye on incorporating the enriched metadata in their centralised indexes

(2) An active approach with regard to the free search engines (Google Scholar, Microsoft Academic Search, Scirus, PubMed and others) to integrate the national link resolver.

(3) An active approach with regard to libraries without link resolvers to use the national link resolvers in any database they subscribe to.

It is foreseen that ABES will evaluate the effects of the steps 1 to 3 on discovery and discovery systems within the French Higher Education community at an appropriate moment.

4.2 Towards a cohesive development for the various components of the national library infrastructure in France

In this paragraph, the cohesion of the development of a number of components of the French national library infrastructure is described:

- **Discovery strategy**: The discovery strategy of ABES is presented in paragraph 4.1 with the development of the national knowledgebase and the national locator service and the Metadata Hub. In this description, the importance of a collaborative effort of French HE libraries with ABES, Couperin and other parties within France is strongly emphasised.
- **ILS in the cloud:** In a parallel project, ABES is working with a group of French HE libraries to develop functional specifications for a shared Integrated Library System in the cloud. This project will result in a tender procedure at the end of 2013. It is foreseen that this tender procedure will result in the selection of an ILS in the cloud and that a small group of libraries will start the migration of the local library system to this ILS in the cloud. It is expected that this migration will take 1 to 1.5 year, so that in the course of 2015 a limited group of French HE libraries (probably 5 to 10) will share an ILS in the cloud.
- **National Licenses:** ABES and INIST are working on the expansion of national licenses and the development of the ISTEX platform for the storage and publication of the content of these licenses. The national licenses are focused on the establishment of retrospective collections of important scholarly content.

In figure 5 the cohesion between these three developments are shown:

- The collaborative effort with regard to the national knowledgebase data collection and the national locator service will precede the ISTEX platform and the ILS in the cloud.
- When the content of the national licenses will be stored on the ISTEX platform, this will mean that the national knowledgebase data with regard to this content has to be adapted.
- When the first group of French HE libraries will have implemented the ILS in the cloud, it is expected that gradually other libraries will follow in this migration to the cloud. From the perspective of SUDOC, the national knowledgebase and the national locator service, this group of HE libraries will be treated in first instance in the same way as the libraries with a local library management system.
- In 2016, there will be a clear perspective on how many and how fast the migration of the library systems to the cloud will take place. Based on this insight, a number of decisions have to be taken *when* and *how* the following migrations can take place:
    - The migration of the union catalogue SUDOC to the cloud, its relation to the shared ILS in the cloud and the development of the mechanism that ensures that libraries with local library systems can take part in SUDOC.
    - The migration of the national knowledgebase to the cloud and become an integral part of the ILS in the cloud in combination with a mechanism that ensures that libraries with local library systems can take part.

- The migration of the national locator service to the cloud and become an integral part of the ILS in the cloud in combination with a mechanism that ensures that libraries with local library systems can use the national locator service.
- At what stage the ILS in the cloud is shared by so many libraries that a webscale discovery service as part of the cloud system can function as a national discovery tool for the French HE community.

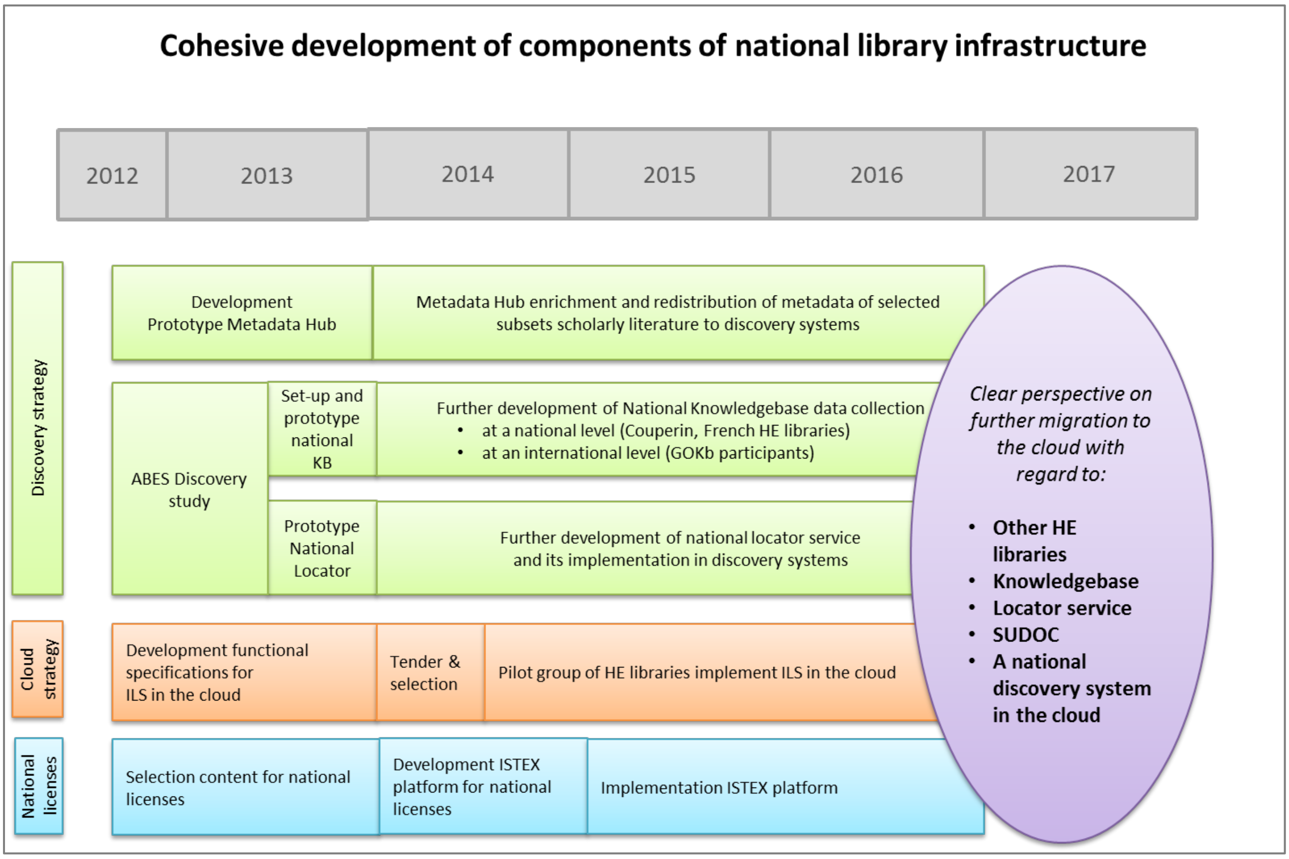

Cohesive development of components of national library infrastructure
2012
2013
2014
2015
2016
2017
Discovery strategy
Cloud strategy
National licenses
Development Prototype Metadata Hub
Metadata Hub enrichment and redistribution of metadata of selected subsets scholarly literature to discovery systems
ABES Discovery study
Set-up and prototype national KB
Further development of National Knowledgebase data collection
• at a national level (Couperin, French HE libraries)
• at an international level (GOKb participants)
Prototype National Locator
Further development of national locator service and its implementation in discovery systems
Development functional specifications for ILS in the cloud
Tender & selection
Pilot group of HE libraries implement ILS in the cloud
Selection content for national licenses
Development ISTEX platform for national licenses
Implementation ISTEX platform
Clear perspective on further migration to the cloud with regard to:
• Other HE libraries
• Knowledgebase
• Locator service
• SUDOC
• A national discovery system in the cloud

Figure 5 Cohesive development of various components of the national library infrastructure

These developments together should deliver a state-of-the-art French national library infrastructure that will improve the information services for the entire French HE community while making the infrastructure more efficient by an optimal sharing of resources by the various partners in the French national library infrastructure.