Project Panorama - vistas on validated information

Eric Sieverts, Marjolein van der Linden, Joost Kircz
Media, Information & Communication Programme
School for Design and Communication
Amsterdam University of Applied Sciences
e.g.sieverts@hva.nl, m.van.der.linden@hva.nl, j.g.kircz@hva.nl

1. Introduction

Internet has become the foremost information source for a large majority of people, young and old, in the Netherlands certainly not less than in other European countries. Search is a ubiquitous functionality available at any website. General search engines like Google, Yahoo! or Bing have become the primary tool for locating information for almost everybody. According to ComScore, in December 2009 every single minute an average of almost 3 million web searches were performed worldwide (1). Furthermore the simple search engine interfaces act as a usability benchmark for any search system. Moreover, web search has made "discovery" and "delivery" of information to coincide (2). What you find (discover) in Google is immediately displayed (delivered) on your screen; a mechanism very unlike classical library practice. As a result a user-expectation of "instant satisfaction" of any information need has developed. Unfortunately, often these facts also trigger the general notion that no other information exists than what can be found by these search engines.

From these developments one might get the impression that the general public has no problems anymore to obtain any information they need. In practice this turns out to be only seemingly true, as is demonstrated by the following observations.

- A growing number of people wonder how reliable the information is, that can be found so easily by Google.
- Many people still cannot find the information they need, despite (or just because of?) the mind-blowing amount of information - many hundreds of billions of web pages and documents - indexed by the spiders of Bing, Google and Yahoo!.
- Most people are unable to filter or refine search results of millions of hits in a reasonable way, and therefore depend on only the first five results, ranked highest on the basis of some mysterious relevancy factors, only known to a few Google or Microsoft engineers.
- Although there are countless, specialized search tools that grant selectively access to merely validated information and can act as alternatives for general search engines, their existence - let alone their coverage - are unknown to most people.
- People having located trustworthy information, published in professional or scholarly magazines that do not support the "open access" paradigm, cannot have access to digital versions of such articles for free (or at a reasonable price), due to license restrictions of commercial publishers, unless they belong to some privileged group of university patrons - as a result "instant satisfaction" is no reality for most of them.

These observations indicate that "access for all" has not yet been realised for all types of information for all types of users. This situation is at variance with Dutch government policy, which pretends to focus on a knowledge economy, an important aspiration of the European Lisbon treaty of 2000 (3).

These considerations were the motivation for Project Panorama, which aims to realise a search system which addresses most of these points. It should
- offer free access to a search system,
- in which a comprehensive selection of validated information can be found,
- by means of a user-friendly one-stop search and find process,
- offer interpretation of value and meaning of retrieved information in its proper context,
- and indicate the best (i.e. cheapest and most convenient) way to obtain full versions of this information - especially for licensed material.

In addition Panorama had a secondary goal, a kind of hidden agenda. Commercial publishers provide digital access to their journal articles on the basis of bulk contracts, often for library consortia. These licences allow access, strictly limited to use within a building or campus, or by well-defined user groups of its own students and staff. Most publishers have no business model yet, what additional rates should be charged to for instance a university library that wants to serve as an intermediary for specific other types of users to provide them access to this licensed material. A side-effect of Panorama would be that the real need for licensed information by non-regular user groups could be better quantified, providing arguments to breach the present vicious circle of discussions on this issue between publishers and libraries.

Even if the main goal of Panorama is described in a single bulleted sentence as in the previous paragraph, it sounds quite ambitious. Therefore, as a preliminary phase of this project, a feasibility study has been performed [(4)]. Its main objectives were:
1. to obtain a more precise understanding of the problem and assess its pertinence;
2. to learn from existing projects and systems with similar goals;
3. to find approaches how to determine what (types of) information to include and how to organise this process;
4. to establish requirements for a search system for this material, that looks as simple to use as Google;
5. to identify possible methods to provide users with licensed information.

In this paper we report the findings of this study, together with some recent developments in relation to Panorama [(5)].

2. Findings

2.1 Expected need and pertinence

To us, as experienced information professionals, the underlying problem was already quite clear. However, discussions about the actual needs, with various types of key stakeholders in the information sector, were much less conclusive. They provided very inconsistent evidence for the expected real need of a single general system like Panorama. For certain audiences, like for instance patient groups, specific resources have already been developed or collected. It was beyond the scope of the present study to explore the user needs and user satisfaction for such specialised systems. One of our contacts expected that people would increasingly prefer social (web 2.0 type) networks to get their information needs satisfied. Another remark was the expectation that technical and scientific information has often to be translated towards the level and the specific goals and context of individual users or user groups.

Despite these dissenting views, a reasonable number of interviewed people saw some role for a Panorama system. Their major remarks about the added value that Panorama should offer, amounted to the following:

- Assistance with filtering and selection of search results is more important than mere search help.
- There exists a wide spectrum of user types with different needs and expectations.
- Only in the field of diseases and health, people often want to know "everything".
- Information must be enriched with additional data for interpretation and translation.
- This project must exploit existing trends for co-operation in the library scene.

Most of these points were in accordance with the ideas that existed already at the start of this project.

While working on the project, a government commissioned study about the future of the Dutch public library sector has been published [(6)]. One of the conclusions of this study was that high priority must be given to the integration of digital information services for the general public. This can also be considered as strong support for the basic concepts behind Panorama.

2.2 Existing systems

During our research we did not come across other projects or existing systems elsewhere, with a scope completely similar to Panorama. By targeting a bit lower, however, we discovered a number of systems which had at least some aspects in common with our ideas for Panorama. Mostly such systems are directed towards specific audiences, cover restricted subject fields or contain specific types of material. Despite some exceptions, a common denominator was that most of these systems used metasearch approaches (we will come back to this technical issue in a later section).

Looking in somewhat more detail, we encountered some interesting approaches:

- A Dutch public library system involved a human answering aspect, i.e. anyone could provide answers to factual questions posted at the system. US experience at Yahoo-Answers revealed quite some drawbacks of this method however [(7)].
- Some systems (a.o. Biznar, http://biznar.com/biznar/) provided clustering of search results, based on a statistical word analysis of these results, a convenient method to refine too large search results. Another system showed a clickable word-cloud based on a similar analysis. This functionality can help to solve the user problem of refining search results, as mentioned in the introduction of our paper.
- In one of the co-operation based systems (http://focuss.info/), collaborators could contribute interesting sites to be indexed, by adding them to a shared social bookmarking account on Delicious (a controlled "crowd-sourced" solution).
- To make a collection of sites searchable in a single system, Google CSE (Custom Search Engine) can be used, a solution also applied in Focuss.info; in this way no metasearch approach is needed.
- The metasearch systems Goshme (deep web search: http://www.goshme.com/) and PurpleSearch (licensed bibliographic databases at Groningen University, NL) contain an automatic recommender service; based on the content of individual user queries, these systems decide which search systems or databases qualify most to answer that query.
- Some systems, all of them for health-related topics, provided indications of level or target audience for retrieved items, or split up their system for separate user groups.

We discovered only two projects which aimed at a comparably broad subject scope as Panorama: ReferenceExtract (http://www.referencextract.org/) and Wikia. Unfortunately, while preparing this paper, we observed that the latter has disappeared completely, whereas the former did not yet start, although a year had passed already.

During our study, we did not see systems that used totally new approaches of how to deliver licensed material to wider user groups. Recently, however, one of the identified systems, DeepDyve (http://www.deepdyve.com/), has announced an interesting new policy, offering licensed material for "rent" against relatively low cost, but for a limited period of time. This sounds comparable to certain services for digital music or movie material, but seems somewhat peculiar for written information.

2.3 Selection process

Determining criteria for collection development is everyday practice for libraries. For the evaluation of web sites and other internet resources, well-established quality assessment criteria exist already. Therefore, we did not examine specific solutions for this aspect of Panorama. Moreover an optimal practical implementation can only be established, after major user groups have been identified. For the organisation of selection processes, we expect that Panorama should take advantage of activities already performed by information specialists at various types of libraries and information centres. Especially co-operation within the public library sector - as also advocated in the government report we mentioned in section 2.1 - and within the scientific library sector can contribute to this selection process. The use of web 2.0 methods to share selected resources (like the Delicious example in the previous section) may offer attractive possibilities.

2.4 Search technicalities

There are two main methods to integrate heterogeneous resources behind a single search interface. A majority of the systems which we came across (section 2.2) applied so-called metasearch or "federated search". Every query which is entered in such a system is distributed (federated) to the individual external search systems which have been incorporated in the metasearch system. Mostly the results that are being received from the various systems are collected and combined behind the scenes, in order to present a uniformly formatted and sometimes even deduplicated result list. To facilitate this mechanism, various standard communication protocols for search systems have been developed (e.g. Z39.50 and SRU). The other method, applied in a smaller portion of the systems, uses a search engine of its own. This search engine indexes all the material of the various resources, in order to allow integrated searching of all of it. That is why this technique is nowadays called "integrated search".

Main advantages of federated search are:
- it uses existing search systems, so that there is no duplication of indexing efforts;
- it is technically easier to implement than integrated search.

There are also a number of disadvantages however:
- it can offer only a common denominator of functionalities of all the systems;
- as a result very limited sophistication of search processes can be realised;
- since it is often impossible (and even unwanted) to send every query to all incorporated systems, for each query a selection must be made to what systems to pass it on;
- response times are typically long, since the system has to wait for the slowest answer to arrive;
- most federated search systems for this type of material are not very user-friendly.

Integrated search does not suffer these limitations. Their opposites are the major advantages of this method. It can provide sophisticated, user-friendly and fast retrieval. In turn it has some disadvantages as well:

- implementing and configuring a search engine is technically speaking more complex (although the open source search engine Lucene/Solr has simplified this process considerably);
- it is often complicated to obtain and locally store all the data (or metadata) to be indexed, or to get guaranteed unlimited access to externally stored licensed data for your spider software, in order to get it indexed.

Many institutions in the Netherlands started several years ago to offer federated search solutions to grant their customers access to the variety of databases they have licences for. Now many of them consider to switch to integrated search, because they appreciate its advantages as more important for their purely scientific information. For almost ten years, the Omega system at Utrecht University demonstrates the feasibility of this approach. In the case of Panorama, neither of the two techniques is expected to offer an optimal solution by its own, because for different types of material different disadvantages prevail. Therefore, most probably a combination of the two must be realised.

2.5 User context

For the delivery phase of licensed digital information, identity management and user authorisation play an important role. Many users belong already to one or more groups that have "certain" rights on "some" licensed content. On the basis of identity management, information can be provided under what conditions a user can get access to a retrieved information item. In case a user has no rights whatsoever - as long as no national licences exist yet - a system must automatically decide what alternatives there are.

A first type of alternative is to determine whether (access to) the information can be located close by in physical sense. Think of a public library or a university at bicycling distance, where the information can be accessed locally, within the building. Something similar exists already for physical objects in Worldcat, where you can enter your postal ZIP-code to identify the nearest libraries having a certain book in their collection. With the advent of GPS localisation and local internet services on mobile phones, many more developments of this kind can be expected.

Another type of alternative is to check what other providers or next-best versions of the selected information exist. Technology for providing such alternatives exists. An example of this technology is SFX, commercially developed by ExLibris. On the basis of standardised metadata, this system first checks whether a user's organisation has a digital license for the publisher of a periodical in which an article has been published. Next it locates alternatives, e.g. access to a copy available from another provider or discovered in Google Scholar, or otherwise it just activates a photocopy request. A clever combination of these existing techniques can already provide most of the functionality required for directing users to the most appropriate way to obtain access to required licensed information.

2.6 Functional Requirements

The basic ideas behind Panorama, together with the technical considerations in the previous two sections, may lead to a preliminary set of functional requirements for the system already.

- The system offers a user-friendly one-stop shopping single search interface on all selected material.
- Material, selected for Panorama, for which no selective search system exists already, should be indexed by Panorama's own search engine.
- Material, selected for Panorama, which can not easily be indexed, but for which a search system exists already, must be incorporated in a metasearch part of the system.
- For any query the system must automatically decide to which search systems it must be federated, its own search engine definitely to be included.
- Results from all metasearch questioned systems and from the own search engine must be merged into a single results list.
- Search results must automatically be clustered into groups, based on similarity of subject or similar context.
- Search results must automatically be divided into facets ("faceted search") on the basis of formal metadata for target audience, level, document type etc.
- Automatic analysis of search results should generate relevant words or concepts which could be appropriate to refine that search, like for instance in the "Aquabrowser" interface (see: http://www.medialab.nl/).
- On the basis of an identity management system, in combination with determination of geographic location, the user should be directed to the most appropriate source for full access to the retrieved information.

For the actual design of a Panorama system, in a later stage much more detailed technical requirements have to be specified on the basis of these points.

3. Conclusions

3.1 Original recommendations

From out original study we could not yet draw definitive conclusions about the feasibility of a complete Panorama system. Technically speaking there are no obstacles to realise a system that satisfies the global requirements as described in section 2.6. Uncertainty about the real needs for such a general system, however, prevented a final decision. This is mostly caused by the general observation that there is probably more need for subject or user-group specific services, than for a one-size-fits-all solution. More comprehensive surveys of potential user groups and interviews are required to obtain a more realistic vision. A preliminary conclusion which could be drawn already, was that a Panorama system can very well act as a backbone infrastructure, invisible for end-users, on the basis of which specific services can be developed.

3.2 Latest developments

After we completed our feasibility study, two important external developments have taken place.

1. The earlier mentioned report on the future of the Dutch public libraries [(6)] has resulted in a restructuring of the overall organisation of this sector. One of the outcomes was the creation of a new institution which is responsible for the digital backbone for public libraries. As a spin-off, a co-operation has developed between this institution,

the Dutch National Library and the Dutch university libraries. This co-operation focuses on an improved national information infrastructure. This involves the creation of a comprehensive union catalogue of the physical collections of these organisations and the availability of an integrated search facility and digital accessibility of scientific and technical journal articles. This last component was an important focus point of Panorama as well.

2. The original initiator of the Panorama project has recently been appointed director of the Dutch National Library, the "Koninklijke Bibliotheek". In its Strategic Plan 2010-2013 (8) which has been published earlier this month (11 January 2010), some important ambitions are specified in connection to the Dutch national information infrastructure. The National Library will - among others - turn into a kind of back-office for the provision of digital publications for an as-wide-as-possible audience. It is the ambition that public, as well as academic libraries can serve their own users on the basis of this national infrastructure. Key in this development is that the National Library will serve as a central licensing contact for commercial publishers, in order to break open the scholarly and professional information market for this wider public.

3.3 Final conclusion

The developments which we just described make it likely that at least part of the objectives of the Panorama Project will be realised in the near future. This being the case, even if no actual Panorama system will be developed. And also even if the Dutch national information infrastructure will initially provide only access to digital versions of classical resource types like books, reports and journal articles, and not yet to purely web based resources (which may provide valuable validated information as well). In this way improved information access for all, as was intended by Panorama, will be greatly promoted.

(1) comScore Reports Global Search Market Growth of 46 Percent in 2009 (http://www.comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009)
(2) Lorcan Dempsey - Librarians and the long tail - D-Lib Magazine, Vol. 12, No. 4, April 2006 (http://dlib.org/dlib/april06/dempsey/04dempsey.html)
(3) See for instance: http://www.europarl.europa.eu/summits/lis1_en.htm
(4) A Dutch language version of the final report of this feasibility study is available at http://sieverts.pbworks.com/f/panorama.pdf
(5) PowerPoint presentation available at: http://www.slideshare.net/sieeg/project-panorama-vistas-on-validated-information
(6) Innovatie met effect - Rapport Adviescommissie Bibliotheekinnovatie (2008) (http://www.minocw.nl/documenten/bieb%20innovatie%20jp.pdf) [only in Dutch]
(7) Jacob Leibenluft - A Librarian's Worst Nightmare - Slate, 7 December 2007 (http://www.slate.com/id/2179393)
(8) KB Strategic Plan 2010-2013 (http://www.kb.nl/bst/beleid/bp/2010/index-en.html)