Closing the gap between OPAC and the web: Rich library catalogs

Dipl.-Inf.wiss. Manfred Hauer M.A. (Germany), Head of AGI – Information Management Consultants

Abstract

From 2002 in the United States and German speaking countries libraries started extending their searchable space by adding table of contents of books (TOC) to the OPAC and integrating full text retrieval engines. In total about 2.5 million TOCs have been captured globally. This is an indexing approach providing additional, more specific subject terms than human indexing with classes and descriptors. Searching is becoming more natural. The usage of this new content is increasing substantially as it is very close to browsing the shelves in a well categorized library - but it is independent of time, place and subject. It fits better into the working situation of modern students and teachers who have to be more and more flexible, quicker in achieving exams results and more focused. As often the scanning of all books or periodicals in full text is in conflict with the copyright law. Capturing the TOC fits within copyright law restrictions within most countries and is a vital step in making catalogs more efficient, transparent and easy to use. Linking to full text or capturing is the next step. An international overview will be given and the methods mainly used in Europe will be presented. The major producer in Europe is the collaborative network of „dandelon.com“. They have generated more than 1 million TOCs in many languages. This solution is able to run OCR (Optical Character Recognition) even for Arabic language publications, which pose serious but solvable challenges. Finally the integration of content into OPACs and search engines will be demonstrated.

1. The need for rich catalogs

In the 1980 when libraries moved from card catalogs to the OPACs most students and scientists did not have computer facilities. These catalogs have received a web interface over the last 10 years but the content has not changed. How people access and use libraries has changed dramatically. Is there any student or scientist at universities of rich countries not using a personal computer of their own for features like communicating, writing papers and presentations, searching for information, doing calculations, visualizations, simulations or programming? Are students in poor countries not interested in having their own computers instead of sharing? There are also many other computer based leisure activities shaping minds and attitudes in the present day.

As soon as one is able to use the features listed above, he or she is able to use the web for publishing documents and research results. Individuals as well as small to very large publishers and content aggregators are using web publishing. So the norm in the mind of clients is „rich and easy available content“. This norm is already true at Open Archives, personal websites of scientists, publishers and some bookseller websites or in the Google Book library. But it is not true for publications printed on paper only, if they are not close at hand. There is a gap and a bridge has to be built.
2. Definition of catalog enrichment

Capturing tables of content started in the US and the German speaking areas in 2002. Some librarians think, the display of TOCs is the major advantage. We think the extension of searchable space is most relevant. First an item has to be found before anything can be read. While the normal OPAC stores on average 5-7 subject words for representing an item of 200 pages, a TOC already enlarges the number of subject words to 50, 100 or maybe 500. The substantive in chapter titles are a strong indexing approach, relevant and precise in most cases. They are selected by the author, the major expert in that item. His language is real world language in that subject area – while the official terminologies of librarians are more abstract and often several years behind real world language development. No doubt, both are useful, but two are stronger than one, if they cooperate. So extending the number of relevant, precise, searchable subject terms makes library searching more „web-like“ and the same is true for displaying it immediately. It gives a feeling close to physically browsing the shelves.

By the way: The extraordinary historical success of Google is closely related to the possibility of searching each word written by the author and by seeing the document without the need to register or go through lending or payment procedures. There is a third factor, the high quality of relevance ranking of search results, which is possible only on rich text collections. Google is the OPAC of the web. While they were just crawling the web in the beginning they turned additionally into a major producer and/or aggregator of content. They are working closer with publishers and authors already as some librarians do. Why not learn from a successful company?

Back to things librarians already do. Capturing only TOCs complies with copyright in most countries as reference information is normally not valuable enough for protection. The same is true for bibliographic information which links to the text but is not able to replace it. It is not the message, just the announcement of the message. It is possible that some local publishers may complain about scanning TOCs, but I am not aware of any juridical problems. The international publishers know this practice and support it. The more searchable their content is the better.

Here is the complete list of approaches for catalog enrichment:

Content – adding more text

<table>
<thead>
<tr>
<th>TOC</th>
<th>As mentioned above, the major approach.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstracts</td>
<td>Partially provided by publishers via ONIX, partially rather „advertising“ style.</td>
</tr>
<tr>
<td></td>
<td>Abstracts are an alternative to TOCs in the case of articles and have been standard in documentation over the last 200 years.</td>
</tr>
<tr>
<td>Part of text</td>
<td>Paragraphs or chapters – copyright restricted.</td>
</tr>
<tr>
<td>Text</td>
<td>Linking or embedding, stronger copyright restrictions, a future main road.</td>
</tr>
<tr>
<td>Reviews</td>
<td>Not available for most books and copyright protected, high quality.</td>
</tr>
</tbody>
</table>
Manfred Hauer: Closing the gap between OPAC and the web: rich library catalogs

**Profile linking**  Linking to “surrounding” information like profiles of authors, research groups, associations, publishers or lists of citations (who read and cited this item).

**Indexing – extracting relevant topics**

*Machne Indexing or text mining* for extracting major subject words and phrases, recognition of people, places or subject classes.

**User Tagging**  Adding descriptors/tags by end users.

**Query expansion**  Executed while doing information retrieval, based on mono- or multilingual thesauri or other language resources.

**Recommender**

*Recommender*  There is a statistical approach monitoring the click activities, this approach is biased to mainstream thinking.

**User rating**  And commenting - a human approach by assessing single items.

**Visual empowering**

**Cover pages**  Nice to have, little relevance for content, easy to manage.

Most of this is implemented at successful players such as Amazon – it has a commercial standard and library users often find books there before they turn to lending.

3. **Method of capturing TOCs**

There are two sources for capturing: firstly digital data – which is true for current digital born books – and secondly the digitization of paper – valid for about 100 million books. As there are a large number of authors and publishers and file formats, handling of digital data is not at all a small task. On the other hand digitization combined with Optical Character Recognition (OCR) is able to rebuild the text most often at a level close to perfect. This is true for about 200 languages. OCR for Arabic, a language without vowels, is more complicated. Two major vendors are competing and large dictionaries for validation are currently under construction.

One *Syndectics* went the first way, collecting digital text from publishers while doing registration for their new books. This is possible, as this company belongs to Bowker, the US ISBN agency. In Germany the other *AGI – Information Management Consultants* built a solution for scanning OCR and machine indexing of TOCs by cooperating and sharing with libraries. Both have generated collections of around 1 million tables of contents. Bowker is renting out that content, while the people capturing through digitization share the results in most cases. In 2003 AGI started a sharing platform called *www.dandelon.com*, they also share with public library service centers. In 2008 TOCs
Manfred Hauer: Closing the gap between OPAC and the web: rich library catalogs

were officially set as a new library standard for Germany and Austria and the German National Library started to generate collections on their own and share it. Liechtenstein National Library became a member of the dandelon.com community as early as 2004. Before this date the Library of Congress started in the US. Switzerland is on the way and there has already been uptake in universities in German speaking areas. Norway has completed an evaluation and may go ahead after having replaced the whole library management system of all public and scientific libraries including the National Library. In other European countries there are small projects as well (Italy, Netherland, Poland), but we are not yet aware of details. The collections generated in the German speaking areas cover titles from more than 100 languages.

We know about 10 different technical solutions. I will focus on the one solution which generated 60 % of the whole collection. It is known as “intelligentCAPTURE mobile”. It’s workflow starts with a barcode reading of the shelf numbers or the imported lists of system numbers. The software is connected to the related library server and retrieves automatically the book details. Protocols used are Z 39.50, SRU or specific XML. If there are ISBN and year available, automatically a lookup for available TOCs at other servers takes place. If the TOC was already generated anywhere else, it is downloaded and the user can store and move to the next book if not, a scanner captures the relevant part of an item. On average 3.5 pages in the case of TOCs. As the library catalog knows the language of an item, OCR and CAI (Computer Aided Indexing – a synonym for text mining) and other optional tasks can run parallel in the background. Quality indicators show which scans should be checked before exporting. The export formats fit the related library management system and can be imported with a scheduled agent. With up to 10,000 books per month at each scanstation the productivity is high. This high speed is possible because of many technical details, but of central importance is the mobility of “intelligentCAPTURE mobile”. The mobile scanstations can even work within narrow book shelves as visible in a video recorded at the State- and University library of Hamburg.

4. Usage and retrieval results: rich versus poor OPACs

In 2008 the German National Library (DNB) started producing and publishing TOCs (Schneider 2009). In 2009 the usage of TOCS increased by 800 %. In 2004 GBV in Göttingen (15 million different books, used in 800 libraries, 2.5 million TOCs) introduced the service and in 2008 they measured an increase in usage of 400 % over and above their increases since 2004 and 2008 (Hauer, Diedrichs 2010). They combine all results of dandelon.com, Syndectics, Library of Congress, German National Library, Casalini and exchange TOCs with other library service centers including the one from Austria. There is no service inside their catalog which has a similar increase in usage – even 6 years after introduction. DNB and GBV started in 2009 with full text search facilities inside the TOCs and relevance ranking. So users are able to find items with highly specific subject terms not searchable with traditional cataloging.

If a library system does not support full text search and relevance ranking (information retrieval) – which is true for many - even with the normal library management system similar effects can be achieved by using text mining. Text mining means the extraction of relevant subject terms, names, places or subject classes and storing it in the catalog records. It can be based more on linguistics or
Manfred Hauer: Closing the gap between OPAC and the web: rich library catalogs

more on statistics. These extracted terms are moved to the normally non-visible part of the record and are searchable.

Can the difference between rich and normal OPAC be measured? In an evaluation project 100 students of information science got an oral description of the major content of 300 different books and had to find these books. They did not know any bibliographic details. It was plain subject search. Libraries of different sizes with and without TOCs were evaluated. First libraries without TOCs are reported: While in very small libraries (less than 5000 books) the books wanted could be found in many cases even without TOCs, in very large libraries (the testing collection was NEBIS from Switzerland with 4.5 million books) students failed in most cases (even if it was their home catalog). At that time only the pioneer in catalog enrichment Vorarlberger Landesbibliothek (400.000 books), had a reasonable collection of searchable TOCs (Raedler 2008). Most books were found with less than 3 searches. In the same study Google Books and dandelon.com was evaluated, dandelon.com had an advantage, however in the meantime the number of items increased significantly, so that the result may not longer be valid (Hauer 2006).

Dandelon.com introduced a search solution in 2004 with relevance ranking, intensive query expansion based on multilingual thesauri, with TOCs, import from abstracts, cover pages, machine indexing and full text search of the TOC, linking to e-books and bookseller and a user interface in 27 languages, including Arabic. The usage is stable but the application did achieve extraordinary success as the same content could be searched and displayed directly in as many as a thousand libraries. The proximity to the physical item matters only as long as the book is not available digitally. With TOCs libraries improved their service and importance.

TOCS uncover the assets of libraries. This step is required before progressing into fully digitally based libraries.

References


Contact