12 (2009)
AUSGABE 2

Reprint

B.I.T.online

Bibliothek Information Technologie

D 52614
ISSN 1435-7607

Zeitschrift für Bibliothek, Information und Technologie mit aktueller Internet-Präsenz: www.b-i-t-online.de

User Report –

C-3 Periodical Server: Automated Cataloguing of Journal Articles

User Report

C-3 Periodical Server: Automated Cataloguing of Journal Articles

Christoph Müller, Rüdiger Stratmann, Nicolai Sternitzke

The library of the Ibero-American Institute (IAI) Prussian Cultural Heritage Foundation in Berlin is Europe's largest special library on Latin America, Spain, Portugal and the Caribbean. It holds 830,000 monographs, 33,000 journals, 73,000 maps and 38,500 audio-visual materials and is the third largest special library of its kind in the world, after the Library of Congress in Washington and the Nettie-Lee-Benson-Collection at the University of Texas at Austin. Each year, 30,000 monographs are added through purchasing, exchange and donations and the library has subscribed to 5,000 journals.

IBERO-AMERIKANISCHES INSTITUT
PREUSSISCHER KULTURBESITZ
IAI
P K

Approximately 65% of the holdings are in exclusive German ownership. The IAI's collection focuses on Humanities, Political Science, Geography, Ethnology, Amerindian Studies as well as Archaeology, each related to Latin America, Spain, Portugal and the Caribbean. In addition, the IAI library oversees the DFG[1] SSG[2] in the subject areas of Latino Studies, as well as Law, Government Publications and Daily Newspapers in or from Latin America.

All materials are made available to users at the institute itself or through the national and international interlibrary loan system as well as the *subito* document delivery service. Every year, more than 15,000 media units are delivered through the various distribution channels.

As a member of the GBV[3] the IAI library uses an OCLC Pica catalogue where the predominant part of the library's holdings is listed. The IAI's collections are administered by the local library system LBS4.[4] Since the foundation of the IAI in 1930, articles from anthologies and journals have been indexed systematically. Since 2000, image files from selected journals have been made available in a current contents database which can be searched by journal title, country of publication, place of publication, subject terms, issue number and call number.

In order to be able to improve the verification situation of journal articles and at the same time the bibliographic supply of information according to the demands of a special library, it was decided in 2006 to officially join the Online Contents Service for Special Collections (OLC-SSG) of the GBV. For this purpose, a technical solution for automated formal cataloguing of the contents of journals was sought. The only half-automatic programme used at that time by different Special Subject Collection Libraries was the *Current Contents Tool* (CC-Tool) which was developed at SUB Göttingen[5]. This tool processes the scanned tables of contents with the OCR-Software *Omnipage* and issues them as ASCII data in a text editor. Title, author and page references have to be marked manually in a highly

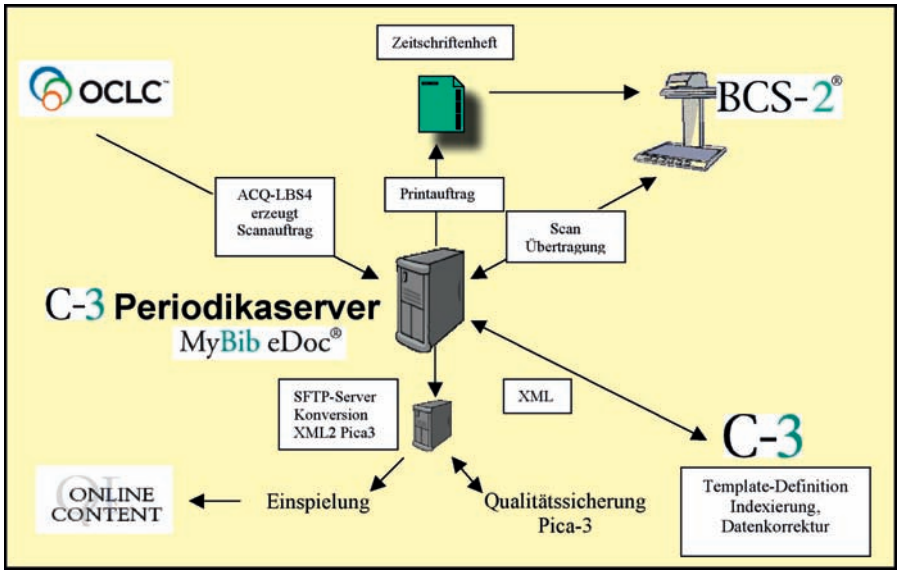
1 DFG = Deutsche Forschungsgemeinschaft (German Research Foundation)

2 SSG = Sondersammelgebiet (Special Subject Collection). The German Research Foundation defines and finances 121 special collections in 40 German Sondersammelgebietsbibliotheken (Special Subject Collection Libaries)

3 GBV = Gemeinsamer Bibliotheksverbund (Common Library Network). The GBV is a common library network of the seven German federal states Bremen, Hamburg, Mecklenburg-Western Pomerania, Lower Saxony, Saxony-Anhalt, Schleswig-Holstein, Thuringia as well as of the Prussian Cultural Heritage Foundation.

4 by OCLC Pica.

5 Staats- und Universitätsbibliothek Göttingen (Goettingen State and University Library), Göttingen, Lower Saxony.



time-consuming process. The result is then converted to PICA3 format by the CC-Tool in a downstream process. This approach, however, appeared to be unsuitable right from the beginning, as the necessary comprehensive editorial work was not in accordance with the demands of the mass cataloguing project planned by the IAI. Moreover, the processing of non-German language content pages with the CC-Tool posed a problem: special characters such as the tildes and diacritics used in Spanish and Portuguese also had to be entered manually – which further made it unsuitable for the library of the IAI. A possible solution for these extended requirements was seen in the C-3 software, which was still in its development stage at that time. The modular software developed by *ImageWare Components GmbH* in Bonn was to enable the automatic recognition of title, author and page reference information in the scanned tables of contents as well as their automatic indexing and conversion into catalogue entries. As pilot libraries, the SUB Göttingen and the ZB MED[6] in Cologne participated in further developing the C-3 programme suite and provided the library-specific requirements for the adaptation of the software to the indexing functionality and a library-related workflow. Since the system's index structure recognition as well as OCR recognition of foreign language texts were not fully mature at that time, the IAI decided to wait until the system had been optimized and to continue using the existing current content service for the time being.

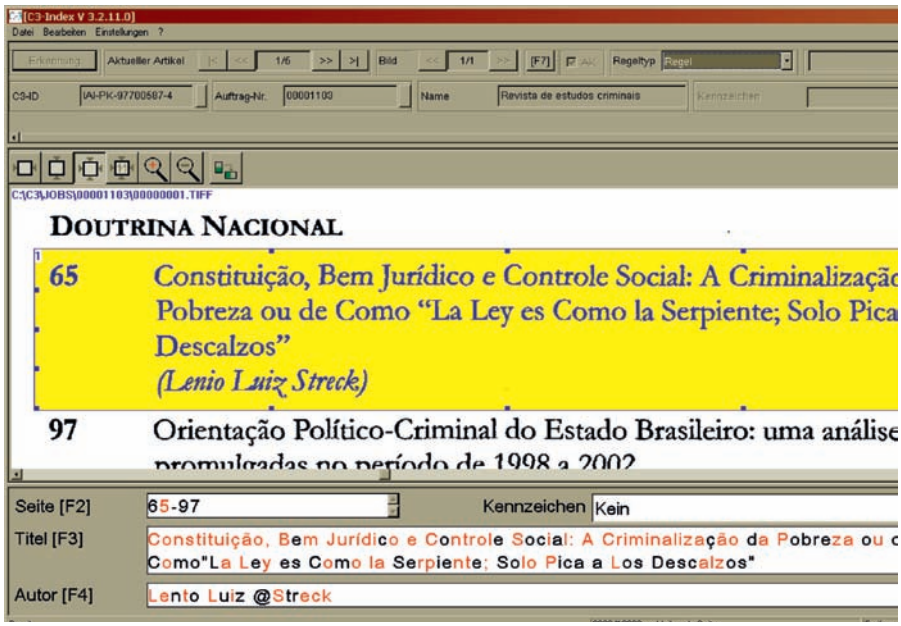
In autumn 2007, when the C-3 software went into productive operation at the SUB Göttingen and the C-3 module had been optimized in fundamental additional functions, the IAI decided to use the C-3 software to index its own journal collections and to make use of the hosting offer *C-3 Periodical Server* of the GBV head office.

Additional workflow components to the C-3 indexing software are the BCS-2® scan client, which is used to scan the journals' content pages that are to be indexed, as well as the MyBib eDoc® system, which is hosted by the GBV head office. It has interfaces to all workflow components, provides a web-based multi user capable order management and controls the workflow with batch processes. In the meantime, special diacritical characters, multilingual contents and even complexly structured contents can be processed without a problem, so that the software makes fully automatic indexing possible.

ImageWare Components GmbH and the GBV head office in Göttingen were commissioned to set up the hardware and software components for two work stations to initiate the scanning and indexing process, one scan station, two C-3 indexing work stations and three SFTP access possibilities for the quality assurance of article data.

As far as hardware is concerned, a *Wide-TEK A3*-Scanner was purchased, which very quickly scans the contents with the specially configured BCS-2 scan software. Otherwise, ordinary standard desktop PCs are used, which provide access to the order management of the central C-3 periodical server via a browser.

The C-3 software, consisting of the *C-3 Template and C-3 index* programme modules, is used at the indexing work stations. The content index structure of each periodical (interpretation type "rule", "table" or "unstructured"), the sequence of publication title, authors and page numbers and the character attributes (bold, italics etc.) of the bibliographical data are once defined by the operator of *C-3 Template*. Based on this template definition, scans of the individual contents are recognized in the C-3 index module with the OCR software *Abbyy Fine Reader,* so that the bibliographic information of the individual articles can be categorized and issued separately per article. The data generated in this process can be corrected, if necessary, with easy-to-use C-3 editing tools and are then exported to the periodical server in XML format. Automatic conversion routines generate article data in Pica3 format from the XML files, which are then imported to the online contents database of the GBV after a final quality control. To ensure that the articles can be unambiguously linked to the journal they belong to, the serial record, the templates and the data of the respective article title are linked via a distinct identity number,



6 ZB Med = Zentralbibliothek für Medizin German National Library of Medicine, Cologne, North-Rhine Westphalia.

the C-3 ID. This number can be the Pica Production Number (PPN) or the so-called "Swets number".[7]

The whole workflow is managed by the MyBib eDoc®-based C-3 periodical server, where the operator can follow the status of the working process at any time. Furthermore, the indexing order is generated on the periodical system and linked to the order metadata, the scanned tables of contents and the XML export file. This way, all data belonging to the order are clearly laid out and comprehensibly managed and can be edited by the operator if and whenever required.

Implementing the automatic initiation to scan and index the journal issue entries in the acquisition module of the LBS4 presented a particular difficulty. The programme, which was developed at the GBV head office for the order generation, was initially only compatible with LBS3 – the previous version of the local library system used at the IAI. Therefore, further developments on the side of GBV head office and interface adjustments by *ImageWare Components GmbH* were necessary, and were already successfully completed in early 2008. The new additional programme reads the bibliographical data from the receipt generated by LBS4, and converts them into an ILL-subito suitable order mail, which in turn generates an indexing order on the C-3 periodical server.

In addition to technical implementation, which lasted from the end of December 2007 to the end of February 2008, also content-related preparations were necessary. The list of journals to be processed in the current contents service of the IAI was put through a content-related and formal review. In this process, the content-related relevance of the journals, the status of the individual subscriptions and the formal suitability of the contents were tested for indexing with C-3. In the end, the list of journals to be processed with the new system included a total of approximately 750 titles.

In addition to these 750 journals, which are successively complemented by new or renewed subscriptions, another approximately 150 titles, which are owned and processed by other libraries but whose content fits in with the journals held by the IAI, were added. On the basis of these 900 periodicals, the OLC SSG *Ibero-America* was generated, which is accessible through the GVK+[8] of the GBV.

After the set-up and consolidation of the technical infrastructure and after the selection of the journal titles to be processed, employees of the IAI were given the necessary knowledge for operating the software components in the framework of a 2-day in-house training session. Directly following this in May 2008, and approximately 6 months after the decision for the C-3 periodical system, the IAI started production. Since then, more than 1,000 article title data sets have been put into GVK+ every month. The quality of recognition – even with complex, multilingual contents – has reached such a high level that quality controls and data corrections need to be performed to a low extent only. Seen altogether, the amount of effort invested on the side of the IAI for the processing of current journal issues through the offered automatic indexing functions from C-3 is less than expected. The generation of orders takes place in the LBS4 and the quality control is done in running operation, i.e. accessioning and catalogue maintenance. Approximately 0,2 FTEs are necessary for scanning and 1 FTE for indexing. The safeguarding of data quality is ensured by a certified librarian at the IAI; all other working steps are carried out by library assistants. Due to trouble-free running of the data production and the comparatively low human resources and expenditure of time, the IAI began the retrospective indexing of journals to be processed for the Online Contents Service of the SSG *Ibero-America* in January 2009. In an initial step, the IAI is indexing all chosen journals published in and after 2000, which was the year the systematic cataloguing of journal articles at IAI was put to a close and replaced by the current contents service. In a second step, all other previous volumes of the journals selected for Online Contents shall be processed. Until the end of March 2009, 850 tables of content from older periodicals could be processed additionally to the normal operating conditions.

On the one hand, given the comparatively high number of differing software components from the point of view of the operator, a consolidation under one uniform user interface and in one integrated programme would be desirable. On the other hand, due to the modular construction, there is the possibility to integrate system components into the local workflow of the IAI library. All in all, the system has proved to be very efficient. Production runs trouble-free and the software is convenient to operate. The high degree of automation is able to quicken and facilitate cataloguing of journal articles for the SSG Online Contents Service.

7 The OLC-SSG system of the GBV is based on the Swets database Online Contents, in which all article data supplied by Swets are listed and can be searched.

8 Another catalogue of the GBV, including, in addition to the regular bibliographical data of the participating libraries, all online contents data of all respective Special Subject Collections.

CONTACT

Dr. Christoph Müller
Rüdiger Stratmann,
Ibero-American Institute
Prussian Cultural Heritage Foundation
Potsdamer Straße 37
D 10785 Berlin
Tel.: +49 (0) 30 / 266 0
Fax: +49 (0) 30 / 266-2503
info@iai.spk-berlin.de
www.iai.spk-berlin.de

Nicolai Sternitzke
ImageWare Components GmbH
Am Hofgarten 20
D 53113 Bonn
Tel.: +49 (0) 228 / 969 85-0
Fax: +49 (0) 228 / 969 85-84
info@imageware.de
www.imageware.de

ImageWare Components GmbH
Am Hofgarten 20
D 53113 Bonn
Tel.: +49 (0) 228 / 969 85-0
Fax: +49 (0) 228 / 969 85-84
info@imageware.de
www.imageware.de