

SUNCAT: the creation, maintenance and challenges of a national union catalogue of serials in the UK

Natasha Aburrow-Jones, Project Officer, SUNCAT.

Introduction

SUNCAT is designed to be the national serials union catalogue for the UK, supplying locations and holdings of serials in research libraries, as well as being a source of high-quality bibliographic records for cataloguers. A "serial", as a definition covers a multitude of publications. AACR2¹ discusses "continuing resources", but a simpler definition is that the term "serial" covers periodicals, newspapers, magazines, journals, annuals, and so forth, which are issued in intervals, whether those are regular or irregular. Serials are prone to title or issuing body changes; mergers; absorbing other titles, and so forth; the cataloguing of serials is rarely straightforward.

This is the first time that such a catalogue has been built in the UK, covering all regions, and subject areas. SUNCAT covers the entire journal as an entity, and does not include article level information; integration into the wider information environment is essential to link SUNCAT into the world of articles. Building SUNCAT has raised some interesting challenges, which are discussed below.

A brief history

In 2001, research was undertaken to examine the feasibility of developing a UK National Union catalogue for all materials². One of the main messages that emerged was a real need for improved information about serials held and their locations and holdings. A major issue raised was that of variable bibliographic and holdings data in library catalogues.

As a result of the UKNUC feasibility study, it was decided to develop a serials union catalogue. Following a

scoping study³, an Invitation to Tender was issued, initially for a two year period, with funding from the JISC (Joint Information Systems Council) and RSLP (Research Support Libraries Programme), with the purpose of building a UK Serials Union Catalogue. (Later funding was supplied by the JISC alone.)

This was scheduled in three phases:

Phase 1, from February 2003 to December 2004, created the catalogue, and populated it with data from the ISSN Register and the CONSER database, along with data from 22 major UK research libraries, including five copyright libraries.

Phase 2, from January 2005 until December 2006, builds on the work undertaken in Phase 1, increasing the number of libraries contributing to SUNCAT. A pilot service was launched in February 2005, which became a full service in August 2006.

Phase 3, running from January 2007, will be a consolidation of the service, including the addition of some unique developments, and resolution of outstanding issues concerned with duplication of records.

The contract for the creation of SUNCAT was awarded to the University of Edinburgh, led by EDINA, one of the two JISC-funded national data centres, based in the University. EDINA provides access to data and research resources to UK Higher and Further Education Institutions. Project Partners to EDINA are Ex Libris, who supply the Library Management System (LMS) that underpins SUNCAT, called Aleph, and have contributed greatly to SUNCAT developments.

Aims

SUNCAT has two primary aims, which resulted from the findings of the feasibility and scoping studies.

For researchers: a source of information about location of serials, including information about access, for print, electronic and any other type of format of serials.

For librarians: a source of high-quality bibliographic records, available for download, enabling libraries to upgrade records on their local catalogues, and to act as a location tool for reference and inter-library loans.

An additional, but important, aim is to raise consciousness of the importance of quality serials information among UK researchers and librarians.

Contributing libraries

SUNCAT could not exist without the hard work put in by its contributing libraries. In Phase 1, the contributing libraries included both national and university libraries, selected on the basis of their significant and large research collections. The CONSER database and the ISSN register were also purchased. The combination of these databases, coupled with the serials data from the 22 Phase 1 contributing libraries, provided over four million bibliographic records to populate SUNCAT, and provide critical mass.

Phase 2 contributing libraries were chosen to increase the number of unique titles in SUNCAT, and thus provide as complete as possible a list of all titles held in the UK. This included data from a further 50 libraries. As well as additional university libraries, some major public

libraries and libraries from learned societies and specialist bodies are included. The inclusion of these smaller, specialist libraries ensures that SUNCAT will hold many unique titles not widely held in the UK, and these collections are made more visible to researchers and librarians alike. Geographically, these libraries cover the length and breadth of the UK – from Stornoway on the Isle of Lewis, south to Exeter, from Belfast in Northern Ireland and the west, to the University of East Anglia in the east⁴.

Technical description

SUNCAT uses the Aleph 500 software, supplied by project partners, Ex Libris⁵. Aleph is used world-wide in many major academic libraries. Aleph possesses the functionality of being able to display information in union view. Essentially, this means that SUNCAT is a physical, centralised union catalogue, with data from all its contributing libraries stored in a single database. Titles are deduplicated to view, so that there is a single bibliographic record for each title with a list of holdings. This deduplication takes place “on the fly”, at the point of display. Records for the same title are matched together in sets, using a complex algorithm, with the fullest record from the set chosen as the one for display. Before load, the SUNCAT Identifier is added. The matching algorithm matches above format, so that records for electronic and print journals will match together, to create a clear and easy display of SUNCAT.

Data manipulation

Processing of the data from contributing libraries has proved to be an interesting experience, showing the range in data quality and local practices, seen not only across libraries using the same LMS, but also those which are library-specific. Harmonisation of data into a form

suitable for loading into the database provides one of the challenges inherent in SUNCAT.

Firstly, the contributing library sends a file of its serials data – bibliographic and associated holdings records – to the SUNCAT ftp server, preferably in a MARC communications format. However, text files, and even word documents and Excel spreadsheets have been accepted. A data specification is then drawn up, based on close inspection of the data, and the answers to two questionnaires supplied by the contributing library. The data is converted only after approval is given by the contributing library; the resulting conversion is checked before the data is loaded.

There is some standard data manipulation, which all files undergo in their data conversion. Some manipulation is Aleph-specific (such as placing the local control number in the 001 tag), and some is more to do with adherence to the rules in AACR2 and MARC (such as changing 245\$h [computer file] to 245\$h [electronic resource]). These are to ensure that these fields – where they exist in the record supplied by the contributing library – adhere to the SUNCAT upgrading standard, which is based on the CONSER minimum standard.

Holdings information is also standardised, as much as is possible. For a coherent display, Aleph uses the 852 tag in union view for both location and the textual summary holdings statement. The contributing library MARC organisational code is placed in the 852\$a; the location is in the 852\$b; the shelfmark in \$h; and holdings information is in the \$3. This is a non-standard use of the 852\$3, but it was the best place for the data to be held, taking into consideration the display in the Aleph OPAC.

As well as this standard manipulation, each library has its own non-standard data manipulation. There are similarities between the data supplied by libraries using the same LMS, but each library has to be treated as unique. These LMS-specific themes are not enough to write a data specification; individual library practices make up the bulk of the data harmonisation. Every library has its own historical practices, and previous LMS legacy issues. These all have to be taken into consideration when writing the data specification. For those libraries which use UKMARC, there is an automatic conversion process into MARC21 (used by Aleph) during the normal conversion routine. For libraries using a non-MARC system, the data has to be placed into MARC21 before the data conversion can occur. This means that a separate specification has to be drawn up, in conjunction with the standard data specification. Fortunately, all the non-MARC libraries in SUNCAT have used a standardised form for their bibliographic records, which means that the conversion into MARC21 has been relatively painless. However, it does mean that the records created tend to be rather brief, and that can cause problems with matching in with records for the same title, thus adding to the duplication of records in SUNCAT⁶.

It is only after thorough checking that the data is considered ready for load. Often, the writing of the data specification has brought some out-of-date practices to the attention of the contributing library, which has led to the bibliographic / holdings records affected being changed as a result.

SUNCAT must maintain the currency of the database. This is achieved by receiving updates from its contributing libraries, which are loaded in at regular intervals.

There are two types of update – a partial file, which contains only those bibliographic and holdings records in the library catalogue that have changed in some way, either new, deleted or altered, and a full file update, which is where the whole serials file is sent for replacing. This latter system of updating records is not without its pitfalls, but is being resolved.

Matching

SUNCAT offers a duplicated union view, so that there is a clear display, with one bibliographic record displaying holdings for all records for that title underneath. In order to achieve this, SUNCAT uses a sophisticated matching algorithm, whereby records for a given title are matched together, but kept separately, and holdings are merged at the point of display.

The algorithm matches records when they are loaded into the database; this algorithm is based on a complex points system, which was originally developed for Melvyl, the union catalogue for the California Digital Library. SUNCAT has made some refinements to this algorithm, to account for varying cataloguing practices. The addition of the BNB and the 7XX added entries (700, 710, 711 tags only) allows a greater flexibility on the matching. A set of records is determined in a three-stage process, involving an initial pool selection to retrieve potential duplicate records, then a quick-match facility (if a record matches with another on title and ISSN, and reaches the number of points over which threshold it is deemed a match, no further matching occurs), and, finally, a full match, if no quick match takes place. This final stage takes into account other, more detailed fields which help identify that particular serial title. Records are merged together to form a set.

The record that is chosen for display is known as the “preferred” record, and is

the fullest bibliographic record in a given set. This is chosen by a points system, governed by a table running behind the scenes. Points are given to the presence of particular fields, such as a 245, an 856, a 110, one point for each 6XX tag. The record in a set with the most points becomes the preferred record. Only the total number of tags is taken into consideration, not the quality of the information therein.

If a set has a title which occurs often to represent different journals, such as “Annual report”, the title will be given fewer points in the matching process than normal, thus ensuring that no quick match will occur. Only a full match is invoked, to ensure that all other data elements are present and match correctly before two such records match together. Such titles are added to a list of common titles, and are an intrinsic part of the matching process. This list does lessen the number of potential mismatches, although, equally, it means that some records will not match when they are supposed to, due to paucity of data in one or both records, and thus adds to duplication.

The matching algorithm matches above format, so that, for example, records for electronic journals and print journals will match together, if they are deemed a match. The decision to do this was based on the premise that it would be of more use to the end user to know that a title existed, in whatever format, through one record, than having multiple records for the same title. This also reduces duplication of records.

One of the major developments that EDINA and Ex Libris have been working on together is the SUNCAT-ID (SUNCAT Identifier). This is an identifier which currently matches existing sets in SUNCAT. In due course, as records are upgraded and

improved, it is expected that one Identifier will represent each title. The development of the SUNCAT-ID entails a major change in the basic concept of the Union Catalogue, which was previously based solely on automatic and dynamic procedures. The ID creates a more “fixed” union catalogue. It is stored in the 049\$a tag, and is applied after data conversion and before data load. Matching has been noticeably improved as a result of the ID, removing overlapping sets (whereby the transitive nature of the matching algorithm means that a record may belong to more than one set). It also means that the database can be maintained more easily, and data improved as records can be merged or separated as necessary, by forcing matches.

Data quality

One of the major issues, first highlighted in the UKNUC report, and confirmed with the building of SUNCAT, is the problem of data quality in both bibliographic and holdings records across all contributing libraries. The use of lower levels of cataloguing standards results in the duplication of journal titles on the database, as there is simply not enough data to match on, so the record will not match in. There are several methods which SUNCAT is utilising to combat this duplication.

Firstly, the matching algorithm has been improved, as discussed above. The inclusion of several new fields and the alteration of the way some extant fields work are designed to improve the matching, and have been relatively successful.

Secondly, the SUNCAT team will be able to match and un-match records from sets, overwriting the SUNCAT-ID with a correct one.

This will be achieved through a function on the Librarians' Interface, discussed below. Database maintenance is an important function of any catalogue, and SUNCAT is no exception.

Finally, contributing libraries will be asked to upgrade any records that they have supplied to SUNCAT which have been chosen as preferred records or have been verified as unique. The upgrading of a record is to the SUNCAT upgrading standard, which is similar to the CONSER minimum standard.

SUNCAT will also continue further investigation into the issues of duplication within the database, maybe through the future use of an API.

SUNCAT is also finding ways of improving the holdings records that are supplied with the bibliographic records. These can be very brief; in order to supply more information, the library name in the holdings is hyperlinked, taking the user to the front page of the library catalogue or website. This will allow the user to repeat the search to gain the most current information. A further development is to link the holdings with the catalogue record on the library's own system for that title; this is being developed at present.

Librarians' Interface

One of the major SUNCAT developments is the Librarians' Interface, which allows any contributing library access to customised reports and download facilities. Access is allowed through an authentication process; for normal searching and viewing, SUNCAT is freely available, and requires no authentication.

Downloading is probably one of the most important functions for a cataloguer. SUNCAT will allow download, after authentication, for all

contributing libraries in a variety of formats. It is anticipated that the most heavily used download format will be that using the z39.50 protocol; however, for those libraries that do not use this, a download from the web function has been developed, which will allow the library to download bibliographic records through a web interface. The record itself will be in a variety of formats, including MARC communications format and text format, specified by the library. All users are able to email a text version of the record to themselves, using a standard save / email function. Assisted matching will help improve the duplication of records in SUNCAT, and allow librarians to verify records as unique. Some records will not match, either because they have no other matching record in SUNCAT, and should not match, or because they do not have enough fields to match with an existing set. For any record that does not match in with any set, a report will be created, customised to be library-specific. The contributing library is asked to check these records through the Librarians' Interface. The record can then be verified as unique; alternatively, it can be matched in with another set, from a selection of records available as part of an automatic process through the Librarians' Interface. The record is downloaded, and added to the library's catalogue; it is also updated in SUNCAT.

The design and functionality of the Librarians' Interface is currently being tested; it is hoped to have it ready for the contributing libraries to use in a few months.

AIMSS

The JISC-funded AIMSS (Automating Ingest of Metadata on Serials Subscriptions) project has been completed. This proof of concept project involved the transmission of

serial holdings information for the Universities of Glasgow and Leeds from Serials Solutions, a Public Access Management Service, to SUNCAT, where the records existing for the universities were updated. ONIX for Serials (Serials Online Holdings) was the format used for transporting the data.

Future of SUNCAT

At the outset, it was known that SUNCAT would face major challenges, not least because of the low and variable quality of data in UK libraries. The data quality has, indeed, been a major challenge, but it is an achievement that the holdings of over 50 major research libraries are on SUNCAT, and these can be viewed through a well-received interface.

The number of duplicated records has been reduced by fine tuning the algorithm and the implementation of the SUNCAT-ID has eliminated overlapping sets. The introduction of a facility to provide manual matching will improve matters; more software solutions are being actively sought to improve duplication still further.

In order to provide an invaluable service to librarians and researchers alike, SUNCAT must fulfil more roles than maintaining its currency. It must provide a stable service, high-quality records for downloading, more contributing libraries, more unique titles, an improved geographic coverage across the UK, linking with related services (such as Zetoc), and different views onto the data (such as an arts-only results screen, or results for libraries within one specific region). Only then will SUNCAT truly be the National Serials Union Catalogue for the United Kingdom.

Based on a presentation delivered at CIG Conference 2006.

turn to **p.6** for further reading and references

Further reading

1. Burnhill, P and Halliday, L (2004). SUNCAT: a modern serials union catalogue for the UK. *Serials*. Vol. 17, no.1 pp. 61-67.
2. Guy, F and Burnhill, P (2005). SUNCAT as a national serials' facility for researchers and librarians. *SCONUL focus*. No.36, pp.36-43.
3. Burnhill, P and Law, D (2005). SUNCAT rising: UK Serials Union Catalogue to assist document access. *Interlending & Document Supply*. Vol. 33 no.4, pp.203-207.

References

1. AACR2, Chapter 12, 12.0A1
2. UKNUC report <http://www.suncat.ac.uk/description/SUNCAT-NUCrep.pdf>
3. Scoping study <http://www.rslp.ac.uk/circs/2002/suncat.pdf>
4. SUNCAT contributing libraries http://www.suncat.ac.uk/description/contributing_libraries.html
5. ExLibris <http://www.exlibrisgroup.com/aleph.htm>
6. SUNCAT technical papers http://www.suncat.ac.uk/librarians/data_processing_initial_load.html ;
<http://www.suncat.ac.uk/librarians/holdings.html>
7. Matching process http://melvyl.cdlib.org/F/9B7H47A3MEPLJS2CDSM7R87N4937ULXA6NS87NTGI1HR1KJ63Y-00097?func=file&file_name=help-expert-merge-cdl90
8. SUNCAT matching algorithm <http://www.suncat.ac.uk/librarians/matching.html>
9. SUNCAT bibliographic standards http://www.suncat.ac.uk/librarians/SUNCAT_bib_standards.pdf
10. The final report submitted to the JISC http://www.jisc.ac.uk/index.cfm?name=project_aimss



Catalogue & Index is electronically published by the Cataloguing and Indexing Group of the Chartered Institute for Library and Information Professionals (CILIP) (Charity No. 313014)

Subscription rates: free to members of CIG; GBP 13.50 for non-members for four issues.

Advertising rates: GBP 70.00 full-page; GBP 40.00 half-page. Prices quoted without VAT.

Submissions: In the first instance, please contact the editor: Penny Robertson, Senior Information Officer, Scottish Library & Information Council, e: scotearl@slainte.org.uk

For book reviews, please contact the Book Reviews Editor: Neil T. Nicholson, Cataloguing & Metadata Services Team Leader, National Library of Scotland, e: n.nicholson@nls.uk

Issue 156 proofread by Cathy Broad

ISSN 0008-7629

CIG website: <http://www.cilip.org.uk/specialinterestgroups/bysubject/cataloguingindexing>

CIG blog: <http://communities.cilip.org.uk/blogs/catalogueandindex/default.aspx>