Metadata Output and Its Impact on the Researcher

Anne Welsh, Lecturer in Library and Information Studies and Programme Director for the MA Library and Information Studies, University College London

Introduction

This article, based on a presentation at the CIG Conference 2014, presents a brief foray into the question of data output and why it may be of interest to the ordinary catalogue user – specifically the ordinary catalogue user who wishes to download a quantity of catalogue data for further analysis at a later date. It highlights ways in which library metadata may be a source for scholarly research as opposed to a simple finding aid and tentatively suggests that library managers, cataloguers and systems librarians might fruitfully test the output options available and compare these outputs with the data available using staff options in-house.

The Purpose of Catalogue Data

Catalogue records as computer data have a comparatively short history. The original focus of the MARC pilot project was "to test the feasibility of a distribution service of centrally produced machine-readable cataloging data" (Avram, 1968), and although Information Retrieval was acknowledged as one of the "criteria to judge the flexibility and usefulness of the format," it was the third in a list of four criteria, top of which was "printing – bibliographic data display in a variety of forms (3x5 catalog cards, book catalogs, bibliographies, etc.)" (Avram, Knapp and Rather, 1968, p. 3-4). Moreover, a limit on the development of the nascent communication format was acknowledged: "Since so little is known about how a bibliographic record will be used in machine-readable form for retrieval, it was only possible to anticipate future applications." (*Ibid*, p. 4).

At base level, then, although MARC was seen as opening doors to new possibilities, at point of original design its main purposes were no different from those set out by Jewett (1852) over a century earlier and highlighted by his report's commissioners: "The most important of them, perhaps, will be the economy of time, labor and expense, required for the preparation of a new edition of a catalogue" (Everett *et al*, 1853, p. x).

Although there has been plenty of research into Information Retrieval from the 1960s onwards, and although MARC has gone through several versions to reach the current MARC 21, it is important to note that the structure of library data, MARC, predates XML, and that structure is not flexible for sharing with other web resources outside libraries, as has been asserted in many other places, from Tennant's (2002) famous 'MARC Must Die' opinion piece through the BIBFRAME Primer (Miller *et al*, 2012) and beyond.

MARC data is not easily compatible with the web, but within the specialist Library Management Systems that have grown up since the 1960s, it provides an environment in which users can search and retrieve surrogate records of a library's holdings, and the geographic spread and long history of the format is such that users can search in a consistent manner across a large number of library databases and collections. Consortia have published catalogue data on the web, requesting their contributing libraries to submit records following the consortium's MARC cataloguing policies (*cf* OCLC, 2008-   ) or creating mapping tables and cross-walks from the originating libraries' own catalogue records to the consortium's own (*cf* COPAC, 2012). These are, clearly, good starting points for researchers, offering a quick way to find a large number of resources that are relevant to a search.

Welsh, A. (2015) Metadata Output and the Researcher, *Catalogue and Index* 178: 2-6.

Catalogue data also continues the tradition of the card and dictionary catalogues in fulfilling Cutter's second object, "To show what the library has" (Cutter, 1891).

User Tasks

Cutter's (1876) Objects, Means and Reasons for Choice had a seminal influence on both the practice of cataloguing and its scholarship. His work placed users at the centre of the cataloguer's concern and identified the ways in which they could search using the technology of the 19th century. His primary object was "To enable a person to find a book of which either (A) the author, (B) the title, or (C) the subject is known." These three entry points (A-C) corresponded to the three common catalogue card runs and dictionary catalogue volumes at the time, and continued into the computer era as three of the main fields by which users could search, in the days before whole-record searching was possible.

These were distinct from his third object, "To assist in the choice of a book (G) as to its edition (bibliographically), [and] (H) as to its character (literary or topical)." These aspects of the record were recorded not for search but so that, having performed a search, the user could differentiate between the records he found. The means for achieving objects (G) and (H) was largely through the use of notes; again a feature that has continued to the present day – the edition itself appearing in its own space in the record (later MARC field 250) but often supplemented by notes, and the item's "character" described in structured or unstructured notes (later MARC 5XX fields).

We can see here that in predicting the ways in which a user might search the catalogue, the 19th century librarian was limited by the technology of the card and dictionary catalogue: it was not until machine-readable cataloguing was relatively well-developed in the 1980s that whole record keyword searching was possible (Bowman, 2007). Being able to search in this way notionally freed librarians and catalogue users from the limitations of the structured search by a limited number of entry points, but it was not until the draft of RDA (Joint Steering Committee for RDA, 2008) that a move away from main and added entries was mooted. Even this has not been brought to fruition: the MARC record is structured around main entry (fields 100, 110, 111, 130) and added entry (fields 700, 710, 711, 730), and so this 19th century concept remains with us.

Before there was RDA, there was FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). As well as giving us the WEMI (Work, Expression, Manifestation, Item) model fundamental to modern cataloguing, FRBR set out the other linchpin in current cataloguing theory, the "four generic user tasks." Reminiscent of Cutter's Objects, they are "to find … to identify … to select … to acquire or obtain. [IFLA's underlining]" When asked in an exam or elsewhere what we are enabling users to do when we catalogue, these, then, are the activities we would naturally discuss and be expected to discuss.

However, just as the FRBR entity relationship diagram is not the whole of WEMI (Coyle, 2014), the four generic user tasks are not the complete user picture. This is made clear in a close reading of FRBR itself: "The tasks are defined in relation to the elementary uses that are made of the data by the user." (IFLA Study Group on Functional Requirements for Bibliographic Records, 1998). The use of the word "elementary" is significant: it acknowledges that these are only a starting point for our understanding of what it is that catalogue users are doing when they are using our catalogues. Until recently, when ethnographic methods for observing “Catalog users ‘in the wild’” have begun to be employed (Wilson, 2015), research into user behaviour has focused almost exclusively on these tasks, which, in turn are focused on search and disambiguation of records (“find” and “identify” in FRBR terms) as a means to the end of obtaining a full-text resource (“select” and “acquire or obtain”).

Catalogue Records as Textual Cultural Artifact

Certainly, the search and inventory functions of the catalogue continue to be its most important facilities both for the library and its users: when we catalogue, we do so to fulfill Ranganathan’s (1931) ‘Five Laws of Library Science’, in the order in which he organized them (his capitals):

> BOOKS ARE FOR USE. (p. 1) …
> EVERY PERSON HIS OR HER BOOK! (p. 75) …
> EVERY BOOK ITS READER. (p. 299) …
> SAVE THE TIME OF THE READER. (p.337) …
> A LIBRARY IS A GROWING ORGANISM. (p. 382).

However, as cataloguing theorists point out, the catalogue is useful as an object of study not solely in terms of its primary intended uses, but also as a cultural artifact (Smiraglia, 2008). As Kate Whaite (2013) has put it, “A catalogue that is in use is a finding tool, but when a newer version is introduced, the old catalogue becomes a relic of its time.” Moreover, the catalogue record constitutes not only paratext for the item which it describes, but also a written text in its own right. This is not only significant, as Andersen (2002) highlights in terms of Information Retrieval – that users who understand the structure and content of the text of the catalogue record are better positioned to interrogate it successfully in their search enquiries. Nor is it only cataloguers and cataloguing theorists who are aware of the catalogue as text and textual artifact.

Significantly, here I want to stress that researchers are beginning to recognize that quantitative tools used in scholarly research might yield interesting findings when applied to catalogue records. Here, library research intersects with the Digital Humanities, applying software tools devised for Computational Linguistics, and repurposing library tools for Collection Management to answer scholarly questions. In this sense, we are beginning to see an answer to Wilson’s (1968) questions about the “exploitative power” researchers bring to the catalogue – using it as a source, in itself, to answer what we may consider to be standard Humanities questions, summarized by Smiraglia (2008) as “the power of a scholar to make the best possible use of recorded knowledge” (p. 35). In Wilson’s (1968) terms, this power was the greater of the “two kinds of power [in] bibliographical control” – greater by far than the general “descriptive power” on which cataloguing and cataloguing theory have been focused, for pragmatic reasons. To borrow and amend Les Coleman’s (2002) famous quote on printing – for us to know the catalogue to be worth studying, it has to be studied.

Quantitative Analyses

Quantitative analysis of literary texts has a long history (Burrows, 1992), dating back to the pre-computer era (Hoover, 2008) and sometimes attracting controversial responses from literary scholars who feel that literature cannot and should not be quantified (Corns, 1991). Recent examples of computational analysis of texts include Gibbs and Cohen’s (2011) reanalysis of Houghton's (1957) *The Victorian Frame of Mind*, which examined the use of emotion-rich words by Victorian authors, to determine the mood of the era; and the use of text-mining techniques to reveal the structure of Pynchon's novel *V* (Tsatsoulis, 2012). Quantitative approaches have also been used in studies of author attribution (Stamatatos, 2009) and, of course, cultural trends (Michel *et al*, 2011).

Within Bibliography, quantitative methods are beginning to lead to significant projects, such as *Early Modern Print*, which "offers a range of tools for the computational exploration and analysis of English print culture before 1700" (Humanities Digital Workshop at Washington University in St Louis, 2013-), including the EEBO-TCP N-grams Browser, and the EEBO-TCP Keywords in Context tool. CERL (the Consortium of European Research Libraries) (2012-…. b) provides the Material Evidence in Incunabula (MEI) service which builds on the records of the Incunabula Short-Title Catalogue, adding notes on manuscript notes and provenance, including links to the CERL Thesaurus of Provenance names and to the CERL Thesaurus of Place Names, which includes geo-coordinates. The Heritage of the Printed Book Book (HPB) Database, also managed by CERL (2012- a), combines data about hand press materials from catalogues across Europe and North America, providing researchers with a single, consolidated file that they can search and / or use for further quantitative analysis.

These CERL projects take as their starting point MARC records, which are then enriched within datasets that we might recognize as scholarly databases, and, while we might be excited by their possibilities, we might also question the limitations of our own, originating library catalogues, and why they do not already contain such enrichment that will be of use to scholars. We might, in effect, question just how far we have come, since Attar's (2004) article eleven years ago asserted "the developing function of a catalogue record as a research tool in itself, instead of a mere finding aid" (p. 11).

Within Digital Literary Studies, there is an interest in computers as 'writing machines' and catalogue output has been identified as a form of computer writing: "Library catalogues over the globe spew out countless replies to queries (author, keyword, call number, title, subject heading, year, language, editor, series)" (Winder, 2008). As cataloguers, we know that the data that is output is written by us – only the order is changed by the means of retrieval. We decide the level of cataloguing we will carry out on each item within a collection; we use our judgment to describe and provide access to each item to the best of our abilities.

Cataloguer Judgment and the Individual Collection

Although touted as something that has been restored to us by the introduction of RDA, "cataloguer judgment" is an important issue with regard to data presentation and quality. It has ever been with us (Welsh and Whaite, 2012). It is cataloguer judgment that decides which notes we will include (Provenance? Binding? Dust jacket?) and, perhaps more significantly, it is cataloguer judgment, along with the judgment of the systems librarian, that decides the classmark or other device we will use to indicate the cohesion of a small collection within a larger one. The ability of the catalogue user to find, for example, all those books once owned by Walter de la Mare, is greatly aided by the use of a separate classmark as well as by a provenance note that informs of the previous ownership. At Senate House Library, for example, it is possible to use an advanced catalogue search on "Mixed classmark" for "WdlM" which retrieves every item from the Walter de la Mare Working Library and the De la Mare Family Archive of Walter de la Mare's Printed Oeuvre. This is significant in allowing the researcher not only to "find … identify … select … [and] obtain" (in FRBR terms) the materials, but to isolate the catalogue records and use them as the basis for her own research database.

If cataloguer judgment is focused solely on the four generic user tasks, it is focused solely on the "elementary" uses that the researcher may make of the catalogue, and this is a missed opportunity for both library and researcher. As Tomm (2012) has pointed out, "The availability of electronic data opens possibilities for general overviews or comparisons now of active interest that were previously either extremely laborious, or simply not feasible, and argues for the great value of consistent forms of description as well as effective access to complete metadata" (p. 72).

In her PhD thesis, Tomm made use of the library catalogue and reference management software in order to manipulate the data about the Raymond Klibansky Collection, which was the focus of her study. The output options from the McGill Library catalogue were not sufficient for her to be able to interrogate the data in all the ways that she needed in order to answer her research questions. James Baker (2013) describes how in a small project to analyse British Cartoon Archive data, he had to run programming scripts to cleanse the library metadata provided to him in XML format, before he could proceed to carry out his quantitative analysis: the data itself in its raw state was not suitable to simply be fed into the software, in this case Voyant tools. In writing about this activity, Baker is far from complaining: he possesses the necessary technical skills to carry out this work with ease, and is sharing his process for the benefit of others with an interest in creating new knowledge from library data.

Similarly, Mitch Fraas (2014) documented how he created a network diagram of Penn codex manuscripts and former owners from MARC data "in the hopes that it will be not only useful to scholars but also might generate some conversation over how libraries and archives distribute their valuable descriptive information." Later in his blog post he asserts, "I realize now that this task would have been near to impossible at most libraries where the online catalogs and back-end databases don't easily allow public users to batch download full records. Fortunately at Penn all of our catalog records are available in MARC-XML form."

In the absence of such data that can be easily downloaded and manipulated, researchers are left to massage catalogue records through one of the outputs provided by the public version of the catalogue. Training for catalogue output focuses on reference management options, which, in turn is focused on keeping track of publications for a bibliography and citing them correctly. Articles that go beyond the simple training of researchers in this basic use focus on tracking where references have been cited (so that, for example, the academic can assess the impact of their published work on later writers).

Perhaps as a result of this focus, I found when I came to download catalogue records for my PhD research that despite an impressive list of export options, there was not a single one that provided me with what I needed: a clean, tab delimited file of MARC fields that I could import into Excel. The CSV and tab delimited text files did not work correctly – even assistance from the then systems team did not result in my having a clean copy of the data. Attempts to export to any of the reference management options did not carry the notes field through, which, given the focus of my work is largely provenance, meant that the most useful elements of the records were lost to me. Inhouse work on the staff version of the catalogue software could have given me a MARC-XML file, or at least a slightly cleaner data dump, and this is the option I would have taken if I were conducting research for an article.

However, PhD research should be one’s own work and it should be possible for others to replicate it – the assumption of the examiners will be that I have obtained the data myself from the public version of the catalogue.

Ahead of the CIG Conference 2014, I tested the output options of the catalogues of other major libraries. I did not find one in which all of the download options offered to the catalogue user resulted in clean records in a format that could be imported to a database or spreadsheet. At the conference itself, I asked for a show of hands from anyone who had tested the output options on the public version of their catalogue. Only one hand, from the British Library, was raised.

The purpose of this paper is not to make us feel bad about our catalogues, or about our lack of prior concern about how out metadata could be used by researchers not just as a source of information about where to find resources, but as research data itself. Instead, its objective is to excite the current cataloguing community with the potential for research that we have in all these catalogue records we have been amassing since the 1960s.

OCLC (©2015) and COPAC (2012-   ) both provide collection management tools that can be used to perform quantitative analysis. At the moment, they are available only to member libraries, and have not been envisaged as tools for researchers. I would argue that they should be; that not only bibliographers and library historians could use them to discover more about particular collections within our great research libraries, but that those with an interest in Big Data could find them useful to add to their equipment for analyzing the huge amount of metadata we have in our libraries. If researchers have been energized to study the mood of the Victorians, surely they can be encouraged to breathe new life into the study of what the Victorians (and others) owned? If digitized books have fed into our quantitative understanding of ‘culturomics’, what do the records of our books’ former owners tell us about cultural trends – in what was kept and valued; how it was dressed up (in bindings); and the extent that ownership was valued (through bookplates, book stamps and simple signatures)?

In reaching beyond our library sphere of data, we don’t simply have to push our data out and link it to the cloud; we can, with a very few checks that our LMS is working, and the sharing of our collection management tools more widely, invite fresh researchers in. As Wilson (1968) asserted in the 1960s, when machine-readable cataloguing was brand new, there are two powers of bibliographic control: descriptive and exploitative. Let us continue in the first of these, as we have proven to be masters of description, while building our skills in the second.

Works Cited

Andersen, J. (2002) ‘Materiality of Work: The Bibliographic Record as Text’. *Cataloging and Classification Quarterly* 33: 39-65.

Attar, K. (2004) ‘Cataloguing Early Children’s Books: Demand, Supply and a Seminar’. *Catalogue and Index* 151: 8-12.

Avram, H.D. (1968) The MARC Pilot Project: Final Report on a Project Sponsored by the Council on Library Resources, Inc. Washington: Library of Congress.

Avram, H.D., Knapp, J.F. and Rather, L.J. (1968) *The MARC II Format: A Communications Format for Bibliographic Data*. Washington: Library of Congress.

Baker, J. (2013) ‘On Metadata and Cartoons’. *British Library Digital Scholarship Blog*, 16 May, http://britishlibrary.typepad.co.uk/digital-scholarship/2013/05/on-metadata-and-cartoons.html

Bowman, J. (2007) ‘OPACS: The Early Years, and User Reactions’. *Library History* 23: 317-329.

Burrows, J.F. (1992) ‘Computers and the Study of Literature’. In C.S. Butler (ed.) *Computers and Written Texts*. Oxford: Blackwell, pp. 167-204.

CERL (2012- a) *The Heritage of the Printed Book Database*, http://www.cerl.org/resources/hpb/main

CERL (2012- b) *Material Evidence in Incunabula*, http://www.cerl.org/resources/mei/main

Coleman, L. (2002) *For It Not to Be Worth the Paper It Is Printed On It Has to Be Printed*. Ballybeg, Tipperary: Coracle.

COPAC (2012) *COPAC: Technical Requirements for New Libraries*. Manchester: COPAC, http://copac.ac.uk/librarians/contributing/Copac%20technical%20criteria%202012.doc

COPAC (2012-) *COPAC Collection Management Tools*, http://ccm.copac.ac.uk/

Corns, T.N. (1991) ‘Computers in the Humanities: Methods and Applications in the Study of English Literature’. *Literary and Linguistic Computing* 6(2): 127-130.

Coyle, K. (2014) ‘FRBR, Twenty Years On’. *Cataloging and Classification Quarterly*, online ahead of print, http://dx.doi.org/10.1080/01639374.2014.943446

Cutter, C.A. (1876) *Rules for a Printed Dictionary Catalog*. Washington: Government Printing Office.

Everett, E., Cogswell, J.G., Folsom, C., Haven, S.F., Hale, E.E. and Livermore, G. (1853) ‘Report of the Commissioners Appointed to Examine the Plan for Forming a General Stereotype Catalogue of Public Libraries in the United States’ In C.C. Jewett. *Smithsonian Report on the Construction of Catalogues of Libraries, and their Publication by Means of Separate, Stereotyped Titles with Rules and Examples.* 2nd ed. Washington: Smithsonian Institute.

Fraas, M. (2014) ‘Charting Former Owners of Penn’s Codex Manuscripts’. *Mapping Books*, 24 January 2014, http://mappingbooks.blogspot.co.uk/2014/01/charting-former-owners-of-penns-codex.html

Gibbs, F.W. and Cohen, D.J. (2011) 'A Conversation with Data: Prospecting Victorian Words and Ideas'. *Victorian Studies* 54(1): 69-77.

Hoover, D.L. (2008) 'Quantitative Analysis and Literary Studies'. In S. Schreibman and R. Siemens (eds.) *A Companion to Digital Literary Studies*. Oxford: Blackwells, http://www.digitalhumanities.org/companionDLS/

Houghton, W.E. (1957) *The Victorian Frame of Mind*. New Haven: Yale University Press.

Humanities Digital Workshop at Washington University in St Louis (2013-) *Early Modern Print: Text Mining Early Printed English*, http://earlyprint.wustl.edu/

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998) *Functional Requirements for Bibliographic Records*. Munich: K.G. Saur.

Joint Steering Committee for RDA (2008) *Resource Description and Access: Draft*. Washington: Library of Congress.

Jewett, C.C. (1852) *Smithsonian Report on the Construction of Catalogues of Libraries and of a General Catalogue and their Publication by Means of Separate, Stereotyped Titles with Rules and Examples*. Washington: Smithsonian Institute.

Meehan, T. (2014) 'What's Wrong with MARC?' *Catalogue and Index* 174, http://www.cilip.org.uk/sites/default/files/Catalogue%20and%20Index%20issue%20174%2C%20March%202014.pdf

Michel, J-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A. and Aiden, E.L. (2011) 'Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* 331: 176-182.

Miller, E., Ogbuji, U., Mueller, V. and MacDougall, K. (2012) *Bibliographic Framework as a Web of Data: Linked Model and Supporting Services.* Washington: Library of Congress, http://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf

OCLC (2008-) *Bibliographic Formats and Standards*, https://www.oclc.org/bibformats/en.html

OCLC (©2015) *Worldshare Collection Evaluation*, http://www.oclc.org/collection-evaluation/features.en.html

Rangananthan, S.R. (1931) *The Five Laws of Library Science*. Madras: Madras Library Association; London: Edward Goldston, http://catalog.hathitrust.org/Record/001661182

Smiraglia, R.P. (2008) 'Rethinking What We Catalog: Documents as Cultural Artifacts'. *Cataloging and Classification Quarterly* 45: 25-37.

Stamatatos, E. (2009) 'A Survey of Modern Authorship Attribution Methods'. *Journal of the American Society for Information Science and Technology* 60(3): 538-556.

Welsh, A. (2015) Metadata Output and the Researcher, *Catalogue and Index* 178: 2-6.

Tennant, R. (2002) ‘MARC Must Die’. *Library Journal,* 15 October.

Tomm, J. (2012) *The Imprint of the Scholar: An Analysis of the Printed Books of McGill University's Raymond Klibansky Collection: A Thesis Submitted to McGill University in Partial Fulfillment of the Requirements of the Degree of Doctor of Philosophy*. Montreal: McGill University, http://oatd.org/oatd/record?record=oai%5C%3Adigitool.library.mcgill.ca%5C%3A114196

Tsatsoulis, C.I. (2012) ‘Unsupervised text mining methods for literature analysis: a case study for Thomas Pynchon's *V*’. *Orbit* 1(2), https://www.pynchon.net/owap/article/view/44

Welsh, A. and Whaite (2012) ‘Our Hybrid History and its Lessons for Today’. *Catalogue and Index* 169: 5-9.

Whaite, K.C. (2013) ‘New Ways of Exploring the Catalogue: Incorporating Text and Culture’. *Information Research* 18(3), http://InformationR.net/ir/18-3/colis/paperS09.html

Wilson, P. (1968) *Two Kinds of Power: An Essay on Bibliographical Control*. Berkeley: University of California Press.

Wilson, V. (2015) ‘Catalog Users “in the Wild’: The Potential of an Ethnographic Approach to Studies of Library Catalogs and their Users”. *Cataloging and Classification Quarterly* 53(2):

Winder, W. (2008) ‘Writing Machines’. In S. Schreibman and R. Siemans (eds.) *A Companion to Digital Literary Studies.* Oxford: Blackwell, http://www.digitalhumanities.org/companionDLS/