Introduction

This methodological communication discusses the use of MarcEdit in a recent research project and foregrounds how a tool designed for the library community to manipulate catalogue data has been repurposed within an academic methodology. As such, it discusses solutions to the research problem generated by difficulties in outputting MARC records highlighted at CIG 2014 (Welsh, 2014) and the IFLA Rare Books and Special Collections Section's Conference *A Common International Standard for Rare Materials: Why? And How?* (Welsh, 2016b) and in articles published in *Catalogue and Index* (Welsh, 2015) and *Cataloging and Classification Quarterly* (Welsh, 2016a). In doing so, it suggests ways in which metadata for a particular set of rare materials – the catalogue records for the Working Library of Walter de la Mare (Senate House Library [WdlM]) – have been incorporated in the research database and thereby moved beyond Wilson's (1968) idea of the "descriptive power" of bibliographic control to the second, greater power he defined – "exploitative power," summarized by Smiraglia (2008, 35) as "the power of a scholar to make the best possible use of recorded knowledge," and which I have previously argued is a larger purpose than those *solely* of applying international standards and creating linked data (Welsh, 2016a).

Marc data as internal library communication

When devised in the 1960s, Machine Readable Cataloging was primarily focused on the solution of contemporary workflow issues in cataloguing. As the Librarian of Congress put it at the time, "The Library of Congress early recognized that the widespread application of computer technology to libraries could come about only if bibliographic data in machine-readable form could be distributed with precision and at reasonable cost" (Mumford, 1968, [i]). Following on from conferences on a machine format for catalogue records, the MARC Pilot Project was established as "an experiment to determine the feasibility of centrally producing a standardized machine-readable record for application by local installations to serve their specific requirements" (Avram, 1968, 9). International interest in the MARC Pilot, and specifically the British National Bibliography's plan for a UK pilot, in the words of MARC architect Avram, "directed thinking toward a standard communications format suitable for interchanging bibliographic data, not only from one organization (LC) to many, but also among organizations, perhaps crossing national boundaries" (Avram, 2003, 1713). The ensuing MARC II Project had four "criteria to judge the flexibility and usefulness of the format," the first of which was "printing – bibliographic data display in a variety of forms (3x5 catalog cards, book catalogs, bibliographies, etc.)," with Information Retrieval appearing third, after "Catalog division – e.g. personal names used as author and subject may be filed together in a separate catalog" (Avram, Knapp and Rather, 1968, 3; underlining in original).

This is not to say that those involved in the initial development of library catalogue data were not interested in how researchers might use their data – merely that anticipating such use was firmly in the area of 'future-casting' for these early developers. As Avram, Knapp and Rather (1968, 4) put it, "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." It is important to acknowledge that the user was at the centre of the motivation to develop machine-readable (and, therefore, machine-retrievable) data. As Byrne (1998, 3) has described, in the pre-MARC environment, "in addition to being focused on a limited number of uses, [1960s – 1990s scientific industry] commercial systems are generally designed for use only by trained personnel, not by the public. Though library systems do include many functions that are designed for use only by trained staff, a primary and essential focus of library systems is use by library patrons rather than trained personnel." She goes on to point out that "Most early library systems, designed in the early to mid-1960s [pre-MARC], were designed to serve as circulation systems. Specifically, these systems were used to maintain records of the items that were checked out by a patron and to produce overdue notices and other forms related to library circulation routines" (Byrne, 1998, 4).

In this, we can see not only that the MARC developers were prescient to consider Information Retrieval, but that they were working in an era in which what and how we catalogued were heavily restricted by the technology available. As Whaite (2013) has argued, "an old catalogue becomes a relic of its time." Considering the date of the MARC reports cited here – 1968 – and of Wilson's seminal *Two Kinds of Power: An Essay on Bibliographical Control* – 1968 – the library catalogues that are "relics of [that] time" reflect the limited technological capabilities of the computer as opposed to the ambitions of developers and cataloguers. That famous year of revolution – 1968 – brought revolutionary changes in the world of library data as well as in society at large. We could suggest that it is important to treat with kindness the restrictions faced by our mid-20th century colleagues, since surely we too have been limited in the late 20th and early 21st centuries by our own computer systems. Surely our own ambitions – for bibliographic models focused on relationships; for data that can be reutilized easily by the world outside the library walls; for improved cataloguing workflows – are not fully reflected in the data we create as "relics of [our] time." There is a time-lag between what we want for our users and its being fully reflected in the data we create.

MARC Data as Textual Artifact

As argued elsewhere (Welsh, 2016a), the narrative arc in the story of the creation of computer catalogue data has, to this point, been focused on library workflows and on search and retrieval. As Byrne and other commentators have described, restrictions in first the hardware and software on which we relied and then the data models available to us have meant that it is comparatively recently – only since the 1990s – that it has been possible to run searches that scan the entire data relating to every item in a library system (*cf.* Byrne, 1998; Bowman, 2007; Tedd, 2007).

Certainly, by the mid-1990s, it was possible for Leeves (1995, 22) to assert, "A common view is that library housekeeping functionality is now well catered for." As library management systems moved from being standalone to networked, it was, arguably, no wonder that cataloguing theorists, seeing records liberated from the local systems of the 1970s and turnkey systems of the 1980s described by Leeves (1995) and Tedd (2007), began to consider new models for catalogue data, such as *Functional Requirements for Bibliographic Records* (FRBR), first published by IFLA in 1998. This model asserts "four generic user tasks … The tasks are defined in relation to elementary uses that are made of the data by the user" and are described as "to find entities that correspond to the user's stated search criteria …; to identify an entity …; to select an entity that is appropriate to the user's needs …; to acquire or obtain access to the entity described" (IFLA, 1999, 82, underlining in original). As highlighted elsewhere (Welsh, 2014; Welsh, 2015; Welsh, 2016a; Welsh, 2016b), it is crucial to recognise that the tasks are "generic" and the uses "elementary" – the IFLA Study Group on Functional Requirements is not saying that these are the *only* uses to which catalogue data may be put. "Search is only elementary" (Welsh, 2016a).

Indeed, catalogue data is more than solely a surrogate for the items it describes; more than simply the objects retrieved in a search. It is, as Whaite (2013) has argued, "a relic of its time"; as Smiraglia (2008) has asserted, a cultural artifact; and as Anderson (2002) has highlighted, a text in its own right. And, as text, it should be possible to interrogate it in the same ways, and using the same tools that we use on any other forms of text, from Pynchon's *V* (Tsatsoulis, 2012) to Houghton's *The Victorian Frame of Mind* (Gibbs and Cohen, 2011) and back into the world of *Early Modern Print* (Humanities Digital Workshop at Washington University in St Louis, 2013-), to name only three projects that have used methods in digital bibliography to explore texts.

MARC data has been used to good effect in the creation of several tools that are of use to bibliographers, including Copac's union catalogue of "c. 90 UK and Irish academic, national and specialist library catalogues" (Copac, 1996-) and Content Management (CCM) Tools (Copac, 2012-); Edina's SUNCAT, the Serials Union Catalogue for the UK (Edina, 2003-); various projects undertaken by OCLC, including its Linked Data Subsets and Linked Data Markup on Worldcat.org (OCLC, ©2016); and CERL's many products, such as Material Evidence in Incunabula (MEI) and the Heritage of the Printed Book (HPB) Database (Consortium of European Research Libraries, 2012-).

However, in terms of study as text in its own right, not solely as information for inclusion in databases for retrieval, attention paid to catalogue data has been scant and difficult. As Mitch Fraas (2014) put it

when describing a project he undertook to create a network analysis of former owners of the codex manuscripts at University of Pennsylvania Libraries, “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.” Similarly, James Baker (2013) has reported a need to thoroughly cleanse data from the British Cartoon Archive before starting the quantitative analysis he wished to carry out on it.

Not all researchers possess the technical skills of Fraas and Baker. Barriers faced include practical issues such as interoperability of tools (Terras *et al*, 2016), and there can be a fundamental lack of understanding of the opportunities afforded by digital research and the datasets that are available (Mahony and Pierazzo, 2012). National libraries have striven in recent years not only to translate MARC data into linked data formats, but also to encourage scholarly engagement with library data through competitions and fellowships (Welsh, 2016a; Welsh, 2016b).

Tomm (2012) reported a methodology that bypassed the need for advanced programming and data manipulation skills:

> Analyses of catalogue records did not proceed directly from the McGill OPAC. Bibliographic data for the Klibansky Collection was exported from the OPAC, gathered in a personal bibliographic database (EndNote) for exploration and manipulation, and then exported again for further manipulation and analysis in a spreadsheet (Excel). Standard desktop software was selected for simplicity and to keep the procedure accessible to a broad group of potential users.
>
> The basic steps are:

1. Export bibliographic data from the catalogue (save .txt file)
2. Import text files of bibliographic data into EndNote
3. Manipulate data in EndNote
4. Export desired fields from EndNote in TAB Delimited [sic] format (save .txt file)
5. Open in Excel for additional manipulation and analysis (Tomm, 2012, 85)

In devising her methodology, Tomm drew on the work of earlier scholars (Gardner et al, 2010; King et al, 2011; Kwan, 2010; Schlichter and Kraemmergaard, 2010; Xu, 2011), although these are focused on the use of reference management software in literature reviews. At the time of writing, Tomm (2012, 77) agreed with Childress (2011, 144) that “Much of the available literature deals largely with citation managers and generators, more specifically reviews, comparisons, evaluations and use cases for such programs.”

However, within the Digital Humanities, we can observe the use of citation manager Zotero for more than solely reference management. The Zotero plug-in Paper Machines has been highlighted as one of the tools powering “The Digital Humanities Contribution to Topic Modelling” (Meeks and Weingart, 2012) and “allow [ing] anyone to begin exploring large collections of sources to look for trends in the data such as an increasing interest in certain subjects over time, which could point the researcher to interesting questions worth pursuing further. With Paper Machines, anyone with a collection of texts stored in Zotero can generate word clouds, phrase nets, map geo-references found in their corpus, extract structured data using DBPedia, or generate and visualize topic models. All of this can be done without having to pre-process your corpus or leave Zotero” (Crymble, 2012). Co-creator of Paper Machines, Jo Guldi has also co-authored *The History Manifesto* (Guldi and Armitage, 2014), which sets out ideas for analyzing the past in order to assist in current political decisions, focusing on the long-term rather than the short-term and therefore on methods which allow historians to conduct such research, inevitably processing Big Data through techniques in Distant Reading. The book includes an explication of the historiographical uses to which Paper Machines can be put:

> One may use the tool to generalize about a wide body of thought – for instance, things historians have said in a particular journal over the last ten years. Or one may visualise libraries against each other – say, novels about nineteenth-century London set against novels about nineteenth-century Paris. Using the tool, a multitude of patterns in text can be rendered visible through a simple graphical interface. Applying Paper Machines to text corpora allows scholars to accumulate hypotheses about *longue-durée* patterns in the influence of ideas, individuals, and professional cohorts.
>
> By measuring trends, ideas, and institutions against each other over time, scholars will be able to take on a much larger body of texts than they normally do. (Guldi and Armitage, 2014, 91).

Even before Paper Machines was developed, Zotero itself was used in projects that chose to visualize their data with other tools such as Voyeur Tools. For example, With Criminal Intent provided a plug-in that enabled users to manage their data from Old Bailey Online through Zotero (Cohen et al, 2011), while Tufts University incorporated Zotero into its Visual Understanding Environment (VUE) alongside other tools (Baepler and Murdoch, 2010, 5). The VUE Project created tutorials on how to use Zotero in 'Dynamic Content Mapping' (VUE Project, 2009b) and 'Semantic Mapping Tools' (VUE Project, 2009a).

Tutorials on Paper Machines include Emory Libraries' *Lincoln's Logarithms: Finding Meaning in Sermons* (Emory Libraries, 2013), although a recent review of *Exploring Big Historical Data: The Historian's Macroscope* (Graham, Milligan and Weingart, 2015) highlights that "the authors had to remove from the book an example using the tool *Paper Machines* because an update to the software on their computers had broken the tutorial in the time it took to write the book," and suggests that "A greater focus on core principles would have been helpful in future-proofing the text" (Crymble, 2016). Book reviewer Crymble's own project, *The Programming Historian*, currently includes three tutorials on "Distant Reading" (Froehlich, 2015; Hulden, 2014; Graham, Weingart and Milligan, 2012), and covers Zotero itself in three tutorials on "Application Programming Interfaces (APIs)" (Morton, 2013a; Morton, 2013b; Roberts, 2013).

Although Tomm did not explore the growing use of Zotero in her 2012 thesis (and some of the more impressive resources relating to it were published after she had submitted her work), its development by Digital Humanists and uptake by Digital Historians further strengthens her assertion that "As catalogues, cataloguing standards and technology outside the library continue to evolve, the ways that catalogue data can be accessed and used by researchers will continue to shift. But a barrier has been broken and catalogue data has already become useful beyond the 'silo' of the library" (Tomm, 2012, 78).

Exporting MARC data from the catalogue

In 2010, I began work on my own PhD, which is an analysis of the books in the Working Library of Walter de la Mare, housed at Senate House Library with the classmark "[WdlM]" (square brackets not indicating an insertion, but punctuation present in the original classmark). The library had recently completed cataloguing materials in [WdlM] and its companion collection of the De La Mare Family Archive of Walter de la Mare's Printed Oeuvre (Classmark [WdlM] T (again, square brackets present in the original)), and the first task I undertook was downloading records to Zotero for my own use.

As reported at CIG 2014 (Welsh, 2014), "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," and I soon discovered "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" (Welsh, 2015, 5).

Although aware of developments around Zotero, my tool of choice for working with catalogue data – especially catalogue data that I need to manipulate – is, of course MarcEdit. As a librarian who qualified in the mid-1990s, who chose to work mainly in small, special libraries and small, special collections within larger

institutions, I'd been using this set of tools, described by creator Terry Reese (2004, 25) as "a free, Windows-based, metadata editing software suite that [he] develop[s] and support[s] as part of [his] contribution to the library profession" almost since its launch in 2000. Unfortunately, and possibly due to the same issues that were affecting the CSV export options, when I imported the data for my PhD to MarcEdit, I received a string of error messages, and my best attempts resulted in data that was messy beyond the limited powers of batch editing at my disposal to fix. From 2010-2014, when I presented my paper at the CIG Conference, it seemed that the only options available to me were to manually tab delimit the hundreds of records (either in notepad or in MarcEdit), or to ask the systems team to create a report just for me using their internal tools or publish the data (again, just for me). As discussed elsewhere, although some libraries are beginning to explore bespoke publishing of data for researchers, there are workflow and cost implications for any one-to-one services (Welsh, 2016a).

Moreover, PhD research should be one's own work, and it should be possible for others to recreate the steps taken to carry it out – from an ethical standpoint it might be improper to ask for bespoke systems work to be carried out on my behalf (Welsh, 2015). Just as Tomm (2012, 85) insisted that "Standard desktop software was selected for simplicity and to keep the procedure accessible to a broad group of potential users," it was important to me that any work to export bibliographic data from the catalogue and import it to my research spreadsheet, or to any other digital tools should be possible to be carried out by me using the tools I had at home – or by my examiners using the same tools, should they choose to check that part of my work.

So I reported to the systems team the issues I had faced as a library user attempting to capture and reuse the data, and I continued with my work, using 'proxy' data created manually, or hacked together as a 'next-best workaround' for a clean dataset. I continued in the belief that over the course of a part-time PhD (5-7 years), something would happen that would make it possible for me to undertake the 'real' work with the 'real' data that I desired. Within my research project, the materials most affected by my use of proxy data were those that don't fall within the areas on which I have focused for full thesis chapters. My work has been structured to consider books that may have had an influence on de la Mare's own writing, such as the poetry collections, short stories and novels he owned, but also those dealing with subjects about which he wrote, including nature, childhood and the supernatural. After an initial inspection of the books in [WdlM] in 2010-2012, my work has been focused on particular genres and subjects within the collection (*cf.* Welsh, 2013a; Welsh, 2013b). So without good-quality data, there might be a danger that materials outwith these areas and whose only appearance may be in the thesis appendices, could be neglected. Books such as Schrödinger's (1944) *What is Life?: The Physical Aspect of the Living Cell* ([WdlM] 488) and Crake's (1874) *Simple Prayers: A Manual of Instruction and Devotion for Schoolboys* ([WdlM] 489) may not take starring roles in the key chapters of the thesis, but they form part of the collection, and should be represented within the data on equal terms with items which are discussed at length.

Following CIG 2014, the idea of exporting data from the catalogue and importing it successfully to appropriate tools grew in importance. A quick assessment of the actual results of data output from a range of catalogues that I had carried out on the build-up to the conference had found similar issues in each that I tried, which had, at first, seemed incredible to me – I thought I must be doing something wrongly. However, when I asked for a show of hands from conference attendees of cataloguers who had tried all the export options available to library users, only one hand was raised – belonging to a colleague from the British Library (Welsh, 2014; Welsh, 2015). Of course, if it were straightforward to export MARC data in easily reusable states, systems teams, library management system vendors, and those publishing datasets in linked data formats would have a much easier working life – and may even find the aspects of their roles that can be described as 'MARC wrangling' would be redundant.

Importing MARC Data to MarcEdit for Mac

In 2015, my belief that "something would happen" that would allow me to export MARC data from the catalogue to MarcEdit was justified. In April 2015, Terry Reese announced that, following demand from the Mac community, he was beginning work on a version of the suite "that uses native Mac APIs" (Reese, 2015). In his own words, "MarcEdit is so fragile when being run on a Mac … [because] MarcEdit utilizes a cross

platform toolset when building the UI which works well on Linux and Windows systems, but tends to be less refined on Mac systems" (Reese, 2015). Interestingly, the reason he gave for not carrying out these developments earlier was lack of perceived demand from Mac users: "I can count on two hands the number of times I've had someone request a version of MarcEdit specifically for a Mac. And since I've been making a Mac App version of MarcEdit available – its use has been fairly low … With an active [MarcEdit] community of over 20,000, I try to put my time where it will make the most impact, and up until a week ago, better support for Mac systems didn't seem to be high on the list" (Reese, 2015). Following a community-led campaign started by Whitni Watkins and Francis Kayiwa, it became clear to Reese that there *was* demand for a stable and reliable MarcEdit for Mac: "After 8 days, it's done. In all, 40 individuals contributed to the campaign, but more importantly to me, I heard directly from around 200+ individuals that were hopeful that this project would proceed" (Reese, 2015).

The Mac Operating System has been gaining in popularity with developers in general. As this year's Stack Overflow survey of over 50,000 developers highlighted, "Last year, Mac edged ahead of the Linuxes as the number 2 operating system among developers. This year it became clear that the trend is real. If OS adoption rates hold steady, by next year's survey fewer than 50% of developers may be using Windows" (Stack Overflow, 2016). From a share of 60.4% in 2013, Windows fell to 54.5% in 2015 and 52.2% in 2016, while Linux use has grown from 19.9% in 2013 to 21.7% this year, and Mac OS X has risen steadily from 18.7% in 2013, to 20.3% in 2014, 21.5% in 2015, and 26.2% in 2016 (Stack Overflow, 2016).

It is not necessary to enter the debates between developers as to which operating system is the best to acknowledge that versions of software for different operating systems can differ in terms of features and efficiency. While MarcEdit, originally created on Windows, was "fragile" on Mac, there are other programs that are created first on Mac and then translated to Windows, and it is not uncommon to run "virtual machine applications" to provide a Mac experience within Windows, or vice versa (Pot, 2016). There are some disadvantages to running one operating system within another, including increased demands on RAM (Random Access Memory) and CPU (Central Processing Unit) use (*cf* Kissell, 2014; Joseph, 2015; Rizzo, 2013). Cataloguing programs Marc Report and RIMMF, designed by The Marc of Quality, as well as Reese's MarcEdit were developed first for Windows. When using them for research, I have relied on a Windows PC at home, while for teaching I have had to use UCL's virtual desktop on my MacBook, which utilises Citrix Receiver to create a Windows environment. The programs do seem to run more slowly on Desktop@UCL than at home on a native Windows interface, and do seem to crash more frequently.

In any case, the development of MarcEdit on native Mac UIs turned out to be the "something" for which I had been waiting. I watched Reese's (2013-) build page until it looked like work was venturing into new territory as opposed to trying to replicate features on MarcEdit for Windows (Reese, 2016), and then in March 2016 I attempted to import data from the Senate House Library catalogue to MarcEdit for Mac – *and succeeded*. Between March and April I played around with the data until I was confident importing it to Excel, firstly to create an overview spreadsheet of all the records in [WdIM], and then to create spreadsheets for specific fields that I could run through Excel and then import to Gephi to create data vizualisations.

In the style and spirit of Tomm (2012, 85) quoted above, here are the steps that I used:

1. Search Senate House Library catalogue using a "Mixed/local classmark" search for "WdlM" (983 records including [WdlM] T and [WdlM] P)
2. Save records 1-557 and export to local disk (export.txt)
3. Open in MarcEditor and check contents
4. Open MarcTools; upload export.txt and execute MARC=>MARCXML function (save as WdlM.xml)
5. Open in MarcEditor and check contents
6. Open MarcTools; upload WdlM.xml and execute MARCXML=>MARC function (save as WdlM.mrc)
7. Open WdlM.mrc in MarcJoin; select "Export Delimited tab"; set delimiter to "Tab(\t)" and check "Normalize data" box. Select desired fields. (save as .txt file)
8. Open Excel; import .txt file (save as .xlsx file).

From this point, I can save .htm files to upload to TAPoR List Words to quantify, for example, the publishers whose books constitute most of the collection. I can create appropriate edges as CSV files to upload to Gephi to create visualizations to help me think about the provenance of the books in [WdlM] – detecting differences in the level of annotation in books written by de la Mare's friends and other books he owned. The .xlsx files generated from export.txt via WdlM.mrc become the central set of data for quantitative analysis.

Being able to work with data from the catalogue itself also allows me to analyse not only what subject headings were used for the books in [WdlM] but also the consistency of their application. Data for [WdlM] and [WdlM] T were not always created 'from scratch' but derived from pre-existing records, and so it is interesting to see how the resulting subject headings were applied. As evinced by Attar's (2012) article on the cataloguing decisions about [WdlM] T, descriptive cataloguing appears to have been a greater focus than subject indexing (Attar, 2012).

Since hearing Caitlin Bailey's (2013) presentation of her MSLIS research into "the use and analysis of small collections in the study of historical thought" and her proposal of "the feasibility of such collections for the development of unique data sets" (Institute of English Studies, 2013), and reading Tomm's quantitative analysis of the subjects within the Klibansky Collection and Davies and Fichtner's (2006) breakdown of the subjects within Freud's Library, I have been keen to analyse (1) the subjects in de la Mare's Working Library (Senate House [WdlM]) and (2) their representation in the subject headings applied to the books. Being able to export data from the catalogue into MarcEdit is the first step in this analysis.

Scope and limitations of use of MarcEdit

Writing in the mid-1980s about the future of research into writers' libraries, Gribben (1986, 311) predicted that "The technology and determination that enable us to penetrate outer space will most likely also give us better means to explore the intellectual lives of our cherished authors. Word-processors, as well as other apparatuses now beyond our ken, will ultimately supplement the researcher's notecards and fileboxes, but an unquenchable curiosity about the creators and backgrounds of great literary manuscripts will continually bring forth dauntless scholars in each generation."

Certainly, reference management software would have been "beyond [Gribben's] ken," far less Williams's 21st century question, "Can we call Zotero a Scholar's Box for the digital age?" and her answer to it, "I think we can, but we need to recognize that the citations we have are still stuck in a box, in many ways. We cannot copy citations from library databases and drop them into a word processor without using a bibliographic manager like Zotero as an intermediary to capture the structured data that might be useful to my computer when I need to format a bibliography" (Williams, 2015,4).

Tomm's (2012, 85) approach to exporting data from the catalogue at McGill had the advantage of using "Standard desktop software … for simplicity and to keep the procedure accessible to a broad group of potential users," but proved difficult to replicate for the Working Library of Walter de la Mare. The tools in MarcEdit ultimately provided not only a way to extract and manipulate data, but to do so in a way that allowed for a great deal of control over the manipulation, since the suite of software developed by Reese allows not only for the execution of a range of different algorithms behind the front end, but also for files to be opened and edited directly through the MarcEditor.

However, in order to use MarcJoin, which proved to be the most useful tool for extracting specific fields, an understanding of MARC 21 fields is required. For example, to extract the dates of publication from WdlM.mrc for use in TAPoR and Gephi, it is necessary to know that MARC 260 $c is the field and subfield for publication date. The availability of the *MARC 21 for Bibliographic Records* online and free of charge means that it would not be impossible for a non-cataloguer to work out the fields they need to create the .txt file they require, but there is, clearly, an extra effort involved. Thinking of Byrne's (1998, 4) differentiation between commercial scientific industry databases and library management systems, we could not claim that MarcEdit has been designed with "a primary and essential focus … [on] use by library patrons rather than trained personnel."

That said, the ability to extract data from the catalogue for use in research by Digital Humanists – whether through the intermediary steps of a reference management system or MarcEdit – can be seen to be the beginning of the fruition of Attar's (2004, 11) prediction of "the developing function of a catalogue record as a research tool in itself, instead of a mere finding aid."

As argued elsewhere (Welsh, 2016a; Welsh, 2016b), the case for the solo researcher working on bibliographic research into an author and / or owner of a private library that has now been absorbed into an institutional library is, in some ways, smaller than the case for the international standardization of rare books cataloguing, or the publication of linked data. However, "In another way, it is much larger, leading us back round to Wilson's philosophy of the exploitative power of bibliographic control. If we can meet the needs of researchers who want to engage with our data not as a route through to 'the real' objects of their research – full-text files, books, the item for which catalog data is a surrogate – but as an integral part of their own research, then, surely, we are assisting not simply in an 'elementary' user task, but something that is fundamental to scholarship: 'the best possible use of recorded knowledge' (Smiraglia, 2008, 35)" (Welsh, 2016a).

To put it in business terms, as Williams (2015, 9) has, "Libraries must find the means by which scholars can save and sort, use and reuse the resources they find from our collections, or faculty will gravitate to for-profit research platforms that will resolve these problems but within a proprietory and private space." More motivationally, "Libraries are part of a generative process. Cards of single ideas are written, rearranged, and stacked to help build theses, which, in turn, help build books which, in turn, form bibliographies, which fill libraries. I'd like libraries to find a way back to Gessner's *Bibliotheca Universalis*, a place where the library and the scholar are connected" (Williams, 2015, 9).

Tools developed in the library domain, such as MarcEdit, provide a connection between my identity as a researcher and my identity as a librarian. I would like to hope that by working on and disseminating methodologies using such tools, the connection between Humanities researchers without a professional library background and the bibliographic data that powers their initial forays into new knowledge can be encouraged and strengthened.

Acknowledgements

Some of the research presented in this article was undertaken as part of a PhD in Cultural Studies at University College London. The author would like to thank Gladstone's Library for the award of a Revd. Dr. Murray MacGregor Scholarship in 2016, which provided a quiet space to write up some of her thesis.

Works cited

Anderson, J. 'Materiality of Works: The Bibliographic Record as Text'. *Cataloging and Classification Quarterly* 33: 39-65.

Attar, K.E. (2012) 'Modern Special Collections Cataloguing: A University of London Case Study'. *Journal of Librarianship and Information Science* 45(2): 168-176.

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

Avram, H.D. (2003) 'Machine-Readable Cataloging (MARC) Program'. In M.A. Drake (ed.) *Encyclopedia of Library and Information Science*. 2nd ed. New York: Marcel Dekker, pp. 1712-1730.

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.

Baepler, P. and Murdoch. C. (2010) 'Academic Analytics and Data Mining in Higher Education'. *International Journal for the Scholarship of Teaching and Learning* 4(2), http://digitalcommons.georgiasouthern.edu/ij-sotl/vol14/iss2/17

Bailey, C. (2013) 'Hidden in Plain View: The Value of the Small Collection'. Presented at: Blackburn's "Worthy Citizen": A Colloquium on the R.E. Hart Collection, London, 23 November.

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.H. (2007) 'OPACs: The Early Years and User Reactions'. *Library History* 23: 317-329.

Byrne, D.J. (1998) *MARC Manual: Understanding and Using MARC Records.* 2nd ed. Westport, Connecticut: Libraries Unlimited.

Childress, D. (2011) 'Citation Tools in Academic Libraries: Best Practices for Reference and Instruction'. *Reference and User Services Quarterly* 51(2): 143-152.

Cohen, D., Gibbs, F., Hitchcock, T., Rockwell, G., Sander, J., Shoemaker, R., Sinclair, S., Takats, S., Turkel, W.J., Briquet, C., McLaughlin, J., Radzikowska, M., Simpson, J. and Uszkalo, K.C. (2011) *Data Mining with Criminal Intent: Final White Paper, August 31, 2011*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.2152&rep=rep1&type=pdf

Consortium of European Research Libraries (2012-) *CERL Resources*, https://www.cerl.org/resources/main

Copac (1996-) *Explore c. 90 UK and Irish academic, national and specialist library catalogues*, http://copac.jisc.ac.uk

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

Crake, A.D. (1874) *Simple Prayers: A Manual of Instruction and Devotion for Schoolboys*. London: A.R. Mowbray.

Crymble, A. (2012) 'Review of Paper Machines, Produced by Chris Johnson-Roberson and Jo Guldi'. *Journal of Digital Humanities* 2(1), http://journalofdigitalhumanities.org/2-1/review-papermachines-by-adam-crymble/

Crymble, A. (2016) '[Review of] *Exploring Big Historical Data: The Historian's Macroscope*'. *Reviews in History*, February, http://www.history.ac.uk/reviews/review/1889

Davies, J.K. and Fichtner, G. (2006) *Freud's Library: A Comprehensive Catalogue*. London: The Freud Museum.

Edina (2003-) *SUNCAT*, http://suncat.ac.uk/search

Emory Libraries (2013) *Lincoln's Logarithms: Finding Meaning in Sermons*, https://disc.library.emory.edu/lincoln/papermachines-2/

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

Froelich, H. (2015) 'Corpus Analysis with Antconc'. *The Programming Historian,* http://programminghistorian.org/lessons/corpus-analysis-with-antconc

Gardner, W.L., Lowe, K.B., Moss, T.W., Mahoney, K.T. and Cogliser, C.C. (2010) 'Scholarly Leadership of the Study of Leadership: A Review of *The Leadership Quarterly*'s Second Decade, 2000-2009'. *The Leadership Quarterly* 21: 922-958.

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

Graham, S., Weingart, S. and Milligan, I. (2012) 'Getting Started with Topic Modelling and MALLET'. *The Programming Historian*, http://programminghistorian.org/lessons/topic-modeling-and-mallet

Graham, S., Milligan, I., Weingart, S. (2015) *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College.

Gribben, A. 'Private Libraries of American Authors: Dispersal, Custody and Description'. *The Journal of Library History* 21(2): 300-310.

Guldi, J. and Armitage, D. (2014) *The History Manifesto*. Cambridge: Cambridge University Press.

Hulden, V. (2014) 'Supervised Classification: The Naïve Bayesian Returns to the Old Bailey'. *The Programming Historian,* http://programminghistorian.org/lessons/naive-bayesian

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.

Institute of English Studies (2013) '[Abstracts for Blackburn's "Worthy Citizen": A Colloquium on the R.E. Hart Collection, London, 23 November]', https://blackhartbooks.files.wordpress.com/2013/06/colloquium-abstracts.pdf

Joseph, C. (2015) 'Parallels, VMware, VirtualBox or Boot Camp: Best Virtualisation Tool for Mac'. *Macworld* 9 October, http://www.macworld.co.uk/feature/mac-software/best-virtualisation-app-run-windows-on-your-mac-boot-camp-vmware-parallels-3626493/

King, R., Hooper, B. and Wood, W. (2011) 'Using Bibliographic Software to Appraise and Code Data in Educational Systematic Review Research'. *Medical Teacher* 33(9): 719-723.

Kissell, J. (2014) 'Running Windows on a Mac: Why I Prefer VMware Fusion'. *Macworld,* 18 June, http://www.macworld.com/article/2364514/running-windows-on-a-mac-why-i-prefer-vmware-fusion.html

Kwan, M.-P. (2010) 'A Century of Method-Oriented Scholarship in the Annals'. *Annals of the Association of American Geographers* 100(5): 1060-1075.

Leeves, J. (1995) 'Library Systems Then and Now'. *VINE* 25(3): 19-23.

Mahony, S. and Pierazzo, E. (2012) 'Teaching Skills or Teaching Methodology'. In B.D. Hirsch (ed.) *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge: Open Book.

Meeks, E. and Weingart, S. (2012) 'The Digital Humanities Contribution to Topic Modelling'. *Journal of Digital Humanities* 2(1), http://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/

Morton, A. (2013a) 'Creating New Items in Zotero'. *The Programming Historian*, http://programminghistorian.org/lessons/creating-new-items-in-zotero

Morton, A. (2013b) 'Intro to the Zotero API'. *The Programming Historian*, http://programminghistorian.org/lessons/intro-to-the-zotero-api

Mumford, L.Q. (1968) 'Foreword'. In H.D. Avram. *The MARC Pilot Project: Final Report on a Project Sponsored by the Council on Library Resources, Inc.* Washington: Library of Congress.

OCLC (©2016) *Innovation Lab: Bridging the Gap between Research and Development,* http://www.oclc.org/research/innovationlab.html

Pot, J. (2016) 'Is One Operating System Not Enough? Here's the Five Best Virtual Machine Applications'. *Digital Trends*, 10 April, http://www.digitaltrends.com/computing/best-virtual-machine-apps-for-mac-linux-and-windows-pcs/#:NoOOLJxH8ABbJA

Reese, T. (2004) 'Information Professionals Stay Free in the MarcEdit Metadata Suite'. *Computers in Libraries* 24(8): 24-28.

Reese, T. (2013-) 'Current News'. *MarcEdit Development*, http://marcedit.reeset.net

Reese, T. (2015) 'Building a Better MarcEdit for Mac Users'. *Terry's Worklog* 11 April, http://blog.reeset.net/archives/1665

Reese, T. (2016) 'MarcEdit Mac: Export Tab Delimited Records'. *Terry's Worklog* 28 March, http://blog.reeset.net/archives/1917

Rizzo, J. (2013) 'How to Run Windows on Macs'. *MacWindows* 23 September, http://www.macwindows.com/winintelmac.html

Roberts, S. (2013) 'Counting Frequencies from Zotero Items'. *The Programming Historian,* http://programminghistorian.org/lessons/counting-frequencies-from-zotero-items

Schlichter, B.R. and Kraemmergaard, P. (2010) 'A Comprehensive Literature Review of the ERP Research Field over a Decade'. *Journal of Enterprise Information Management* 23(4): 486-520.

Schrödinger, E. *What is Life?: The Physical Aspect of the Living Cell.* Cambridge: Cambridge University Press.

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

Stack Overflow (2016) *Developer Survey Results 2016*, http://stackoverflow.com/research/developer-survey-2016

Stalker, L. and Dooley, J.M. (1992) 'Descriptive Cataloging and Rare Books'. *Rare Books and Manuscripts Librarianship* 7(1): 7-23.

Tedd, L. (2007) 'Library Management Systems in the UK: 1960s – 1980s'. *Library History* 23: 301-316.

Terras, M., Baker, J., Hetherington, J., Beavan, D., Welsh, A., O'Neill, H., Finley, W., Duke-Williams, O. and Farquar, A. (2016) 'Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing and Transforming Access to British Library Digital Collections'. Presented at: Digital Humanities 2016, Krakow, Poland, 11-16 July.

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

VUE Project (2009a) 'Map Based Searching and Semantic Analysis Screencast', https://www.youtube.com/watch?v=iKuBE-J7JuU

VUE Project (2009b) 'Mapping a Zotero Collection into VUE', https://www.youtube.com/watch?v=UsYDOo95ses

Welsh, A. (2013a) '"Books! –": *Pleasures and Speculations* in Walter de la Mare's Library'. Presented at: Researching the Reading Experience, Oslo, Norway, 10-12 June, http://discovery.ucl.ac.uk/1413960/

Welsh, A. (2013b) 'The Poet's Poets: Collections and Anthologies in Walter de la Mare's Working Library'. Presented at: Writers and their Libraries, London, 15-16 March, http://discovery.ucl.ac.uk/1414072/

Welsh, A. (2014) 'Metadata Output and its Impact on the Researcher'. Presented at: Cilip Cataloguing and Indexing Group Conference, Canterbury, Kent, 8-10 September 2014, http://discovery.ucl.ac.uk/1450941/

Welsh, A. (2015) 'Metadata Output and the Researcher'. *Catalogue and Index* 178: 2-8.

Welsh, A. (2016a) 'The Rare Books Catalog and the Scholarly Database'. *Cataloging and Classification Quarterly* (in press).

Welsh, A. (2016b) 'The Rare Books Catalogue as the Foundation of the Scholarly Database'. Presented at: IFLA Rare Books and Special Collections Section Conference, A Common International Standard for Rare Materials: Why? And How?, Lisbon, Portugal, 22 February 2016, http://www.ifla.org/files/assets/rare-books-and-manuscripts/Lisbon-2016-presentations/annewelshiflarbsc2016_reduced.pdf

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

Williams, M. (2015) 'Library of Cards: Reconnecting the Scholar and the Library'. *Partnership: The Canadian Journal of Library and Information Practice and Research* 10(2): 1-10.

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

Xu, F. (2011) 'A Standard Procedure for Bradford Analysis and its Application to the Periodical Literature in Systems Librarianship'. *Library Hi Tech* 29(4): 751-763.