ISSN 2399-9667

Catalogue and Index

Periodical of the Cataloguing and Indexing Group, a Special Interest Group of CILIP, the Library and Information Association

Editorial

June 2020, Issue 199

Welcome to issue 199, which focuses on the transformation of data.

We have a great selection of articles that range across MarcEdit, Open Refine, Wikidata and Linked Data, and include some historical overviews to put the discussions into context.

Our first article by Richard Wallis looks back over 25 years of library data and the development of Linked Data, encompassing schema.org and BIBFRAME. He ends with details about an exciting new initiative.

We move on to Alexandra De Pretto who discusses the multiculturalism of metadata; the work that takes place on data being translated between different systems, and how this can be improved by the embracing of new efficient system-neutral models. There is encouragement for libraries to move beyond their outdated systems and towards new efficiencies.

Jason Evans, the first Wikimedian at the National Library of Wales, talks about the work they have done converting library catalogue data into linked open data on Wikidata. They have also contributed 20,000 images to Wikipedia and Wikimedia Commons. His article gives examples of the potential uses of the data and images shared, and the benefits of linked data.

Heather Rosie looks at the metadata behind EThOS, the British Library's e-thesis service, with the interaction between MARC and XML – taking in Qualified Dublin Core, MarcEdit, and MARC Report along the way.

Continuing with the topic of MarcEdit, Concetta La Spada talks us through data manipulation using this tool, and cites real life examples from her work at Cambridge University Press.

Contents

3-8: Getting [library data] from there to here… by Richard Wallis

9-16: The multi-culturalism of metadata, by Alexandra De Pretto

17-21: Library data as linked open data, by Jason Evans

22-26: MARC transformed: MARC and XML – the perfect partnership? by Heather Rosie

27-33: Data manipulation using MarcEdit, by Concetta La Spada

34-37: How Library Carpentry helps us to transform our data and our training, by Phil Reed

38–39: Book review

The final article by Phil Reed exhorts the benefits of Library Carpentry workshops in contributing to a community of practice. Based at the University of Manchester, he has taken part in delivering training to library staff in Manchester and Durham, and is hoping to develop more workshops in the future.

We hope you will find these papers informative and that they will be of benefit to your practice. We would love to hear your opinions, so feel free to contact us.

Philip Keates: p.keates@kingston.ac.uk

Karen F. Pierce: PierceKF@Cardiff.ac.uk

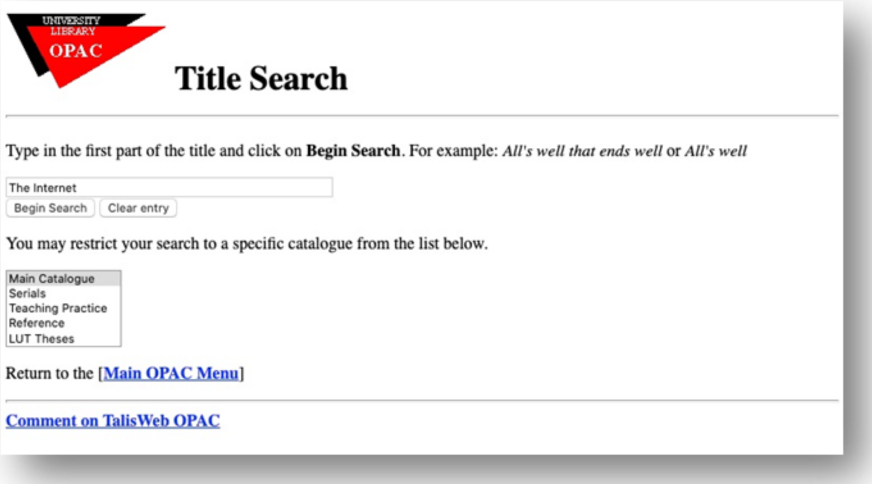
Getting [library data] from there to here…

Richard Wallis, Founder, Data Liberate (http://dataliberate.com)

There are several *here*s and *there*s I will touch upon in this article, but the first that comes to mind is in the East Midlands of England, *circa* 1995.

We are there in the Pilkington Library Building of Loughborough University for the launching of the first Web OPAC. The first installation of a TalisWeb library discovery interface was the result of cooperation between the BLCMP[1] and the library's technical staff.

Previously, staff and students wanting to search the library's online catalogue (OPAC) had to go to the reading room and use one of the dedicated text-based terminals. A technically astute few could potentially also gain access using a Telnet client on the University network.



Title Search

Type in the first part of the title and click on **Begin Search**. For example: *All's well that ends well* or *All's well*

The Internet

You may restrict your search to a specific catalogue from the list below.

Return to the [Main OPAC Menu]

Comment on TalisWeb OPAC

Figure 1: Pilkington Library OPAC c.1999

The mid-90s was the time of the first explosion of the web. The early emergence of the World Wide Web, in the UK university sector, was greatly due to the establishment of the Joint Academic Network (JANET), placing it at the forefront of wider internet adoption. Building on these developments, the team I was working with at BLCMP delivered a web-based window into the library's catalogue. The screen shot of its simple interface reproduced here (with thanks to the *Internet Archive*) is from 1999. Try a comparison with the catalogue discovery interface from your favourite library today.

[1] Birmingham Libraries Co-operative Mechanisation Project – Co-operative and forerunner of Talis Group Ltd

Today’s interfaces are far richer and more colourful. Images feature greatly, layouts adapt to the size and shape of the screen, and links abound to other resources. We have come a long way in the last 25 years. Or have we?

The functionally offered by the vast majority of library discovery interfaces of today is little more than the (web) window into the catalogue that TalisWeb delivered in 1995.

In contrast there have been dramatic developments in the commercial world over the last 25 years – purchasing options directly from search engines; cross-site aggregation of similar results; authoritative information featured on results pages, and increasingly in the voices of personal assistants – all embedded in a plethora of smart devices from tablets and phones to TVs and fridges, to name but a few.

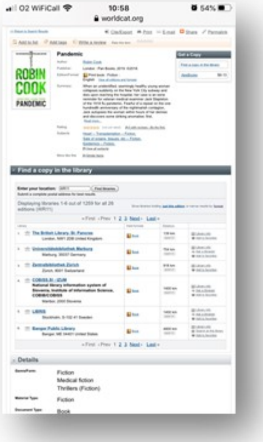

O2 WiFiCall
10:58
54%
worldcat.org
ROBIN COOK
PANDEMIC
Pandemic
Get a Copy
Find a copy in the library
Displaying libraries 1-6 out of 1259 for all 28 editions
The British Library, St. Pancras
Universitätsbibliothek Marburg
Zentralbibliothek Zürich
COBISS.SI - IZUM
National library information system of Slovenia, Institute of Information Science, COBIB/COBISS
LIBRIS
Bangor Public Library
Details
Fiction
Medical fiction
Thrillers (Fiction)
Fiction
Book

Figure 2: Modern OPAC on a mobile

Enabling all this is a detailed understanding, by the search and social networking platform providers, as to what the product, service, or information on a page is about. Increasingly this understanding is being driven by structured data embedded in those pages. Data shared by the sites in structured form (using de facto standards such as Schema.org) are easily understood by the platforms. The consumption of their (freely shared) data helps them become an integral part of these platforms and the web in general.

With a few exceptions, libraries do not share their data in this way. This is why most library web interfaces remain a window from the web into the catalogue, rather than being a route to the web for their rich authoritative data and links to the resources they describe.

Rolling the clock forward a decade to the mid-noughties finds Linked Data[2] on the rise as the way to openly share data across the web. Introduced by Tim Berners-Lee as a pragmatic application of his Semantic Web vision, Linked Data is a recipe – using http protocols, the Resource Description Framework (RDF) metadata modelling framework, and various serialisations such as RDF/XML[3] – to publish data in a consistent, machine-readable way on the web.

For those in the library world experimenting with Linked Data, there was a significant initial data modelling hurdle to overcome. As a sector, libraries were by then well versed in machine-readable metadata. The MARC standard was introduced in the late 1960s, and by the turn of the 21st century the majority of libraries were running sophisticated cataloguing, acquisition, and circulation systems with MARC records at their core.

As a machine-readable representation of the traditional library record index card, a MARC record contains an independent combination of mostly textual information about titles, author and publisher names, subjects etc. Linked Data on the other hand, building on its Semantic Web principles, introduces a different approach by describing entities and their relationships. In a Linked Data dataset, you will find separate individual descriptions of entities such as people, organisations, works, subjects, places, etc., linked together with relationships such as “author”, and “publisher”. All the data is there, but in very different model forms.

[2] Linked Data – Best practices for publishing structured data on the web – https://www.w3.org/wiki/LinkedData

[3] RDF/XML – Serialisation of RDF data using XML

Several groups, including metadata specialists in the British, German, French, Spanish, and other national libraries were up for the challenge. Taking some guidance from established and emerging library standards such as Functional Requirements for Bibliographic Records (FRBR) and Resource Description and Access (RDA), they each developed their own approach to identifying and extracting entity descriptions from their record-based collections to create Linked Data catalogues. These advancements drove a period of enthusiasm, in the second half of the noughties, for the potential for Linked Data in the future for library systems. The British Library Linked Data Model is one example:

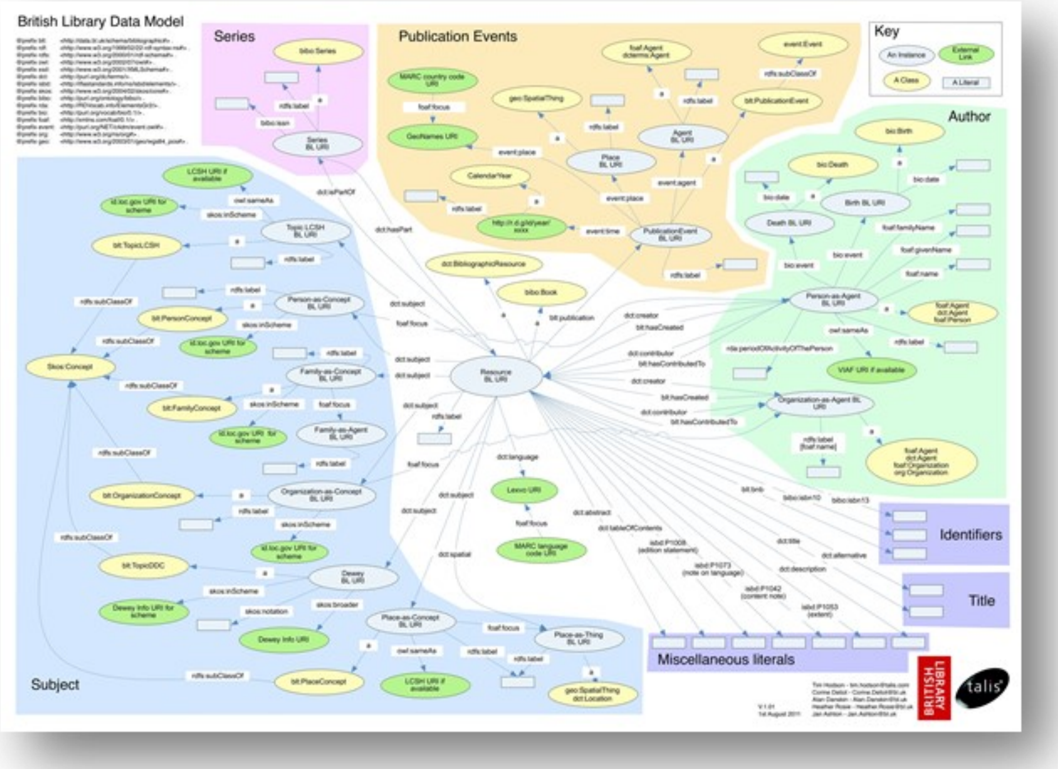


Figure 3: British Library Data Model

Innovations including work- and entity-based navigation in user interfaces, shared open authoritative descriptions for major entities such as people and organisations, linking to external, enriching independent resources, and easier integration with non-library resources were promoted as part of this future vision. The evangelism for this way forward, some of which I happily own up to enthusiastically preaching, majored on it being a way of liberating data from the confines of the library system. The rich data about bibliographic and other resources, painstakingly and expensively catalogued by libraries over many years, trapped *there* in individual libraries, shared out to users and other libraries and thence on to the *here* and now of their users' daily activities.

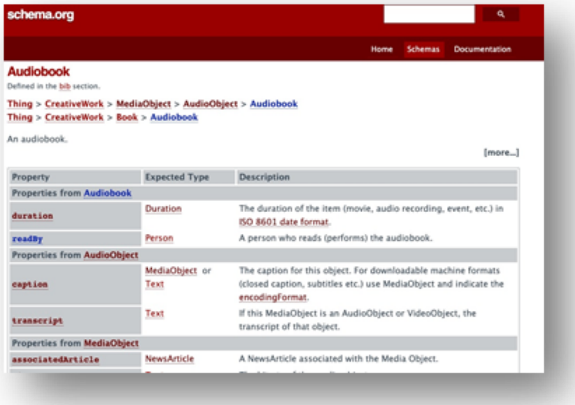
By the beginning of the second decade of the 21st century we were in a great position. Linked data was a hot topic in the library world. No self-respecting library conference could have a programme without one or more sessions with Linked Data in the title, podcasts and blogposts on the subject were being published almost every week, OCLC and other authority sites such as the Virtual International Authority File (VIAF) were publishing Linked Data, and Linked Data principles were starting to influence the development of library standards, for example RDA.

To me, in the heart of that evangelism, however, it was becoming increasingly clear that the momentum for broad adoption of linked data in library systems was beginning to stall.

By then we had some great examples of possibilities – from the British National Bibliography and the Bibliothèque national de France, to name but two. A significant number of bibliographic data sets from many libraries were appearing in The Linked Open Data Cloud. However, there was no significant interaction between the data published by different libraries. No two major implementors of library Linked Data were using the same algorithms to extract bibliographic entities from their records, nor were they using the same combination of Linked Data ontologies, or vocabularies, to describe those entities. As a Linked Data library, you could easily obtain openly shared data from your peers, but interpreting it, and potentially mapping it to, or merging it with, your own data was a significant challenge. Potentially even more significant was the fact that no major library system supplier at that time was seriously considering the adoption of linked data in their systems.

So how have we moved from *there*, in that time of stalling momentum, to *here*, with a rebirth of optimism for a way forward (more of which later)?

2011/12 brought two major developments around standards, one from the Library of Congress, and one from an open community of search engine companies (Google, Bing Yahoo!, and Yandex). The introduction and adoption of these standards have laid the foundations for significant advancements in data sharing that are only now come coming to fruition.



Audiobook

Defined in the bib section.

Thing > CreativeWork > MediaObject > AudioObject > Audiobook
Thing > CreativeWork > Book > Audiobook

An audiobook.

[more...]

Property	Expected Type	Description
Properties from Audiobook		
duration	Duration	The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format.
readBy	Person	A person who reads (performs) the audiobook.
Properties from AudioObject		
caption	MediaObject or Text	The caption for this object. For downloadable machine formats (closed caption, subtitles etc.) use MediaObject and indicate the encodingFormat.
transcript	Text	If this MediaObject is an AudioObject or VideoObject, the transcript of that object.
Properties from MediaObject		
associatedArticle	NewsArticle	A NewsArticle associated with the Media Object.

When first introduced by the search engine community, the Schema.org vocabulary was only seen as a simple set of mark-up tags for embedding into the source of web pages to enable the enhanced display of search results – “Rich Snippets”. Over the last nine years, this structured data, as it has become known, has become a foundation for web sites and pages to share detailed information for search engines to far better understand what they are describing. These structured data techniques are now becoming ubiquitous across the web, and a major part of the search engine optimisation (SEO) community’s tool set. On a topical note, in late March 2020 the Schema.org vocabulary introduced new terms to help governments and companies share definitive data and information about changes introduced in response to the coronavirus pandemic.

Although described as a structured data vocabulary, Schema.org is soundly based upon linked data standards, built upon, and therefore compatible with, Linked Data tools and processes that have gone before. The major step forward it introduced was recommended data serialisation methods (Microdata, RDFa, JSON-LD). These facilitate easy publishing of data within html webpage markup, negating the need for specialist Linked Data servers to share data.

At a similar time, the Library of Congress introduced BIBFRAME, a Linked Data vocabulary and data model, for describing bibliographic resources. It differed from its predecessors, as it had the objective "*to replace the MARC standards, and use linked data principles to make bibliographic data more useful both within and outside the library community*". It also has the influence of the Library of Congress, and its major role in the MARC standards, behind it. As a potential replacement for MARC, used by the vast majority of libraries around the world, it soon gained traction. The *potential replacement* objective was also a disadvantage. To many, its subsequent accommodations to the way things were previously achieved in MARC resulted in it not being the simplest vocabulary to understand from a linked data point of view.

By 2016, the Library of Congress introduced BIBFRAME 2.0, which addressed many of these issues. It also established recommended mappings, and example software, to produce BIBFRAME data from a catalogue of MARC records. Adoption has followed by community groups, such as those behind the Annual BIBFRAME Workshop in Europe and the Linked Data for Production (LD4P) group of major Universities, plus engagement from other key organisations such as the Library of Congress, OCLC, and backing from the Andrew W. Mellon Foundation.

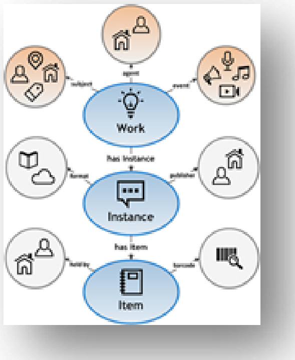


Figure 4: BIBFRAME 2.0 Data Model

It is clear today that BIBFRAME is becoming a de facto standard for bibliographic linked data in the library community. It is criticised both by users of MARC, as not being as flexible, and by linked data protagonists, as being over-complex and too closely associated with its MARC heritage. Yet, most agree that it will suffice for now. Whereas a decade ago no library system developers were thinking about introducing linked data, most will now admit that a minimum capability of importing and exporting BIBFRAME needs to be on their roadmap, and many are considering going further.

This progress stimulated by BIBFRAME is all very well. It doesn't, however, help achieve the goal of getting data out of library-focused systems, to help staff, students, and the general public discover the resources they need. Not in the way that Schema.org does for the commercial and wider web world.

There have been successful experiments in publishing Schema.org structured data from library discovery interfaces. In 2012 OCLC started embedding Schema.org into the millions of pages in WorldCat, a feature that continues today. More recently, the National Library Board of Singapore introduced a set of static web pages to mirror, and link to, its (OPAC) catalogue pages. These Schema.org enriched pages attract significant traffic from search engines and thence on to the catalogue itself.

However, holding back Schema.org development in libraries is the old problem of how to get from record-based data to the entity-based descriptions in an agreed, standard way. So, what is going to get us from *here* to the *there* of libraries being part of the web platform, like the rest of the commercial world? We have most of the building blocks in place.

I would like to introduce an initiative that will bridge these recent developments and enable a way forward. Bibframe2Schema.org is a W3C community initiative, of which I am Chair, to create reference mappings and software from BIFRAME 2.0 to Schema.org. It builds upon the entity extraction and standard Linked Data description established by BIBFRAME to provide a de facto approach to publishing bibliographic data in a form that is consumable by search engines.

bibframe2schema.org

By building on the shoulders of what has gone before, it will mean that, as libraries continue on the road to introduce BIBFRAME-based Linked Data catalogues, they will only have to take a simple step further to expose the benefits of all that effort to the wider population, making their resources discoverable to all who could benefit from them. No longer just a window for people to search through, but part of the normal, data-enabled web. We are not *there* yet, but the stepping-stones from *here* are already becoming visible.

The multiculturalism of metadata

Alexandra De Pretto, Data and Systems Librarian, National Library of Scotland

If ever I am asked to explain in layman's terms what my job is about, the first word that comes to mind with regards to the data aspect is "translation". Indeed, the more I consider it, the more analogies I can make between the experience of navigating between different languages and the work I do with library metadata. Over the course of my life, at different stages and at different levels, I have had the opportunity to interact with seven languages. Being confronted by linguistic differences brings challenges and opportunities. I am an "in-between", which doesn't sound very glamorous, but it makes you pretty polyvalent.

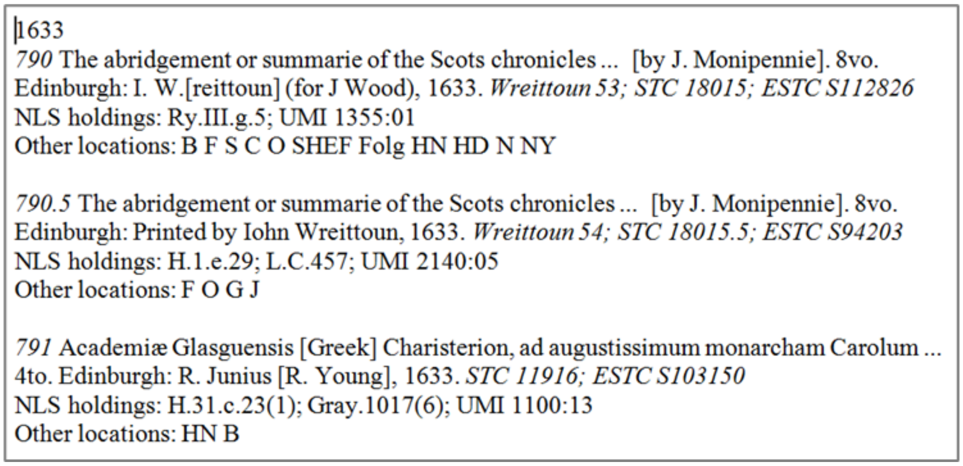
As a data and systems librarian I am no more trained to catalogue than I am capable of mastering seven distinct languages, but I can translate from one to another, in line with my knowledge and ability. This has a lot of advantages and I would like to use this analogy to tell you something of my daily work as Data and Systems Librarian at the National Library of Scotland. In spite of not being a trained cataloguer, CILIP's Metadata and Discovery Group – until recently the Cataloguing and Indexing Group – is the most fitting special interest group for my skills, role and interests, and as such I have been a committee member of the Cataloguing and Indexing Group in Scotland for the past three years.

All of the data and systems work I do tends towards the same aim, which is to improve the accessibility and discoverability of the library's collections: I do this by making the data more consistent, transforming it, or moving it around. It is an indirect form of communication, which is most effective by listening to users' needs and keeping aware of developments in existing and emerging standards, in the same way as one keeps the knowledge of a language alive by practicing and learning.

The diversity of the National Library of Scotland's collections means that my work varies from working with traditional metadata formats to more modern ones, which I will illustrate with a few examples. Ultimately however, such 'translation services' as I provide are not sustainable; the data exchange environment is fast evolving and libraries need to consider how to best adapt by looking beyond current standards and practices, and ask themselves whether dealing with so many 'languages' within a common cultural context is really the best way forward.

From unstructured to structured data

An example of such data translation work is my work on the National Bibliography of Scotland. One of the National Library of Scotland's earliest bibliographies is "*A list of books printed in Scotland before 1701*" by H.G. Aldis, first published in 1904.[1] This resource is in text format, with 6,682 entries currently recorded in a Word document and made available on our website in html format. In other words, this listing doesn't provide much flexibility in manipulating the content; no standard is followed beyond that of the bibliographic transcription and the data is quite sparse:



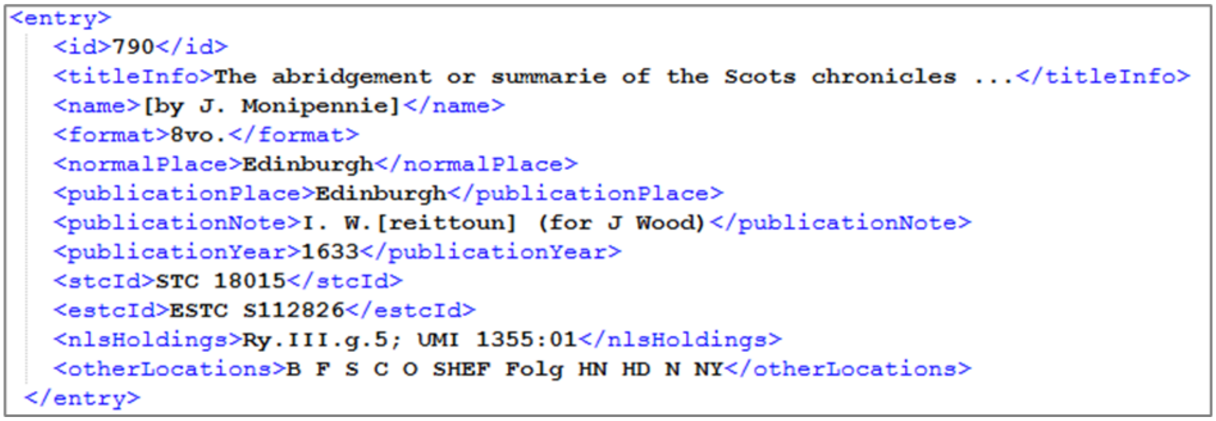
1633
790 The abridgement or summarie of the Scots chronicles ... [by J. Monipennie]. 8vo.
Edinburgh: I. W.[reittoun] (for J Wood), 1633. *Wreittoun 53; STC 18015; ESTC S112826*
NLS holdings: Ry.III.g.5; UMI 1355:01
Other locations: B F S C O SHEF Folg HN HD N NY

790.5 The abridgement or summarie of the Scots chronicles ... [by J. Monipennie]. 8vo.
Edinburgh: Printed by Iohn Wreittoun, 1633. *Wreittoun 54; STC 18015.5; ESTC S94203*
NLS holdings: H.1.e.29; L.C.457; UMI 2140:05
Other locations: F O G J

791 Academiæ Glasguensis [Greek] Charisterion, ad augustissimum monarcham Carolum ...
4to. Edinburgh: R. Junius [R. Young], 1633. *STC 11916; ESTC S103150*
NLS holdings: H.31.c.23(1); Gray.1017(6); UMI 1100:13
Other locations: HN B

Figure 1: Sample of the Aldis Word document listing

'Translating' the standard form of entry in the bibliography starts with the conversion of Word into XML in order to have a more flexible format to work with. Patterns in the entries – words that occur regularly, brackets, dots, digits, or carriage returns – can be analysed and inconsistencies corrected to allow the various entries to be broken down into individual elements. The outcome is a listing of entries with added context:



```
<entry>
   <id>790</id>
   <titleInfo>The abridgement or summarie of the Scots chronicles ...</titleInfo>
   <name>[by J. Monipennie]</name>
   <format>8vo.</format>
   <normalPlace>Edinburgh</normalPlace>
   <publicationPlace>Edinburgh</publicationPlace>
   <publicationNote>I. W.[reittoun] (for J Wood)</publicationNote>
   <publicationYear>1633</publicationYear>
   <stcId>STC 18015</stcId>
   <estcId>ESTC S112826</estcId>
   <nlsHoldings>Ry.III.g.5; UMI 1355:01</nlsHoldings>
   <otherLocations>B F S C O SHEF Folg HN HD N NY</otherLocations>
 </entry>
```

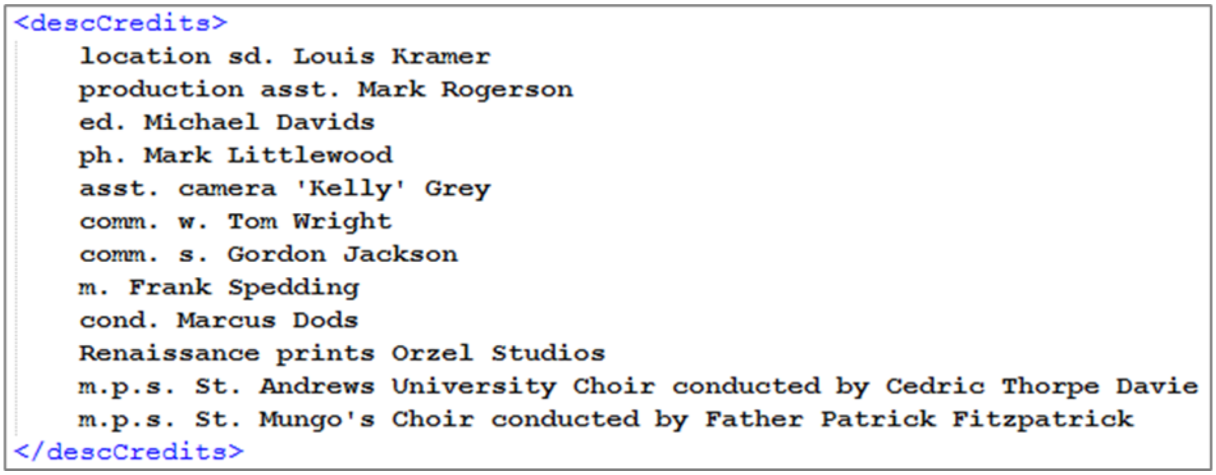
Figure 2: More structure added and splitting of the various data elements

[1] More information about the Aldis list can be found at: https://www.nls.uk/catalogues/scottish-books-1505-1700/introduction [accessed 20/02/2020]

A large proportion of entries have ESTC reference numbers that, once isolated, can be match to the English Short Title Catalogue in order to obtain complete records. Other entries require additional manual involvement but with the result as above, it is possible to produce the skeleton of a record. I rely on cataloguing colleagues – the language experts – to correct mistakes and fine-tune the information extracted through those automated processes. The aim is to have MARC 21 records that can be loaded in our library management system in order to populate an integrated national bibliography for Scotland.[2]

A local idiom

For historical reasons, the National Library of Scotland has developed in-house schemas for specific collections, one of them being the Moving Image Archive (MIA) collection. This data needs to be manipulated in order to be sent to the European Film Gateway (EFG), a portal for historical film archives. Whilst the Library follows the International Federation of Film Archives (FIAF) guidelines in how to describe films, some fields contain a lot of unstructured data and require a high level of interpretation. This is notably the case for the credits field, in which roles and names are recorded. It is a rich source of data, but extracting and analysing the content of this field is particularly challenging. Here is an example:



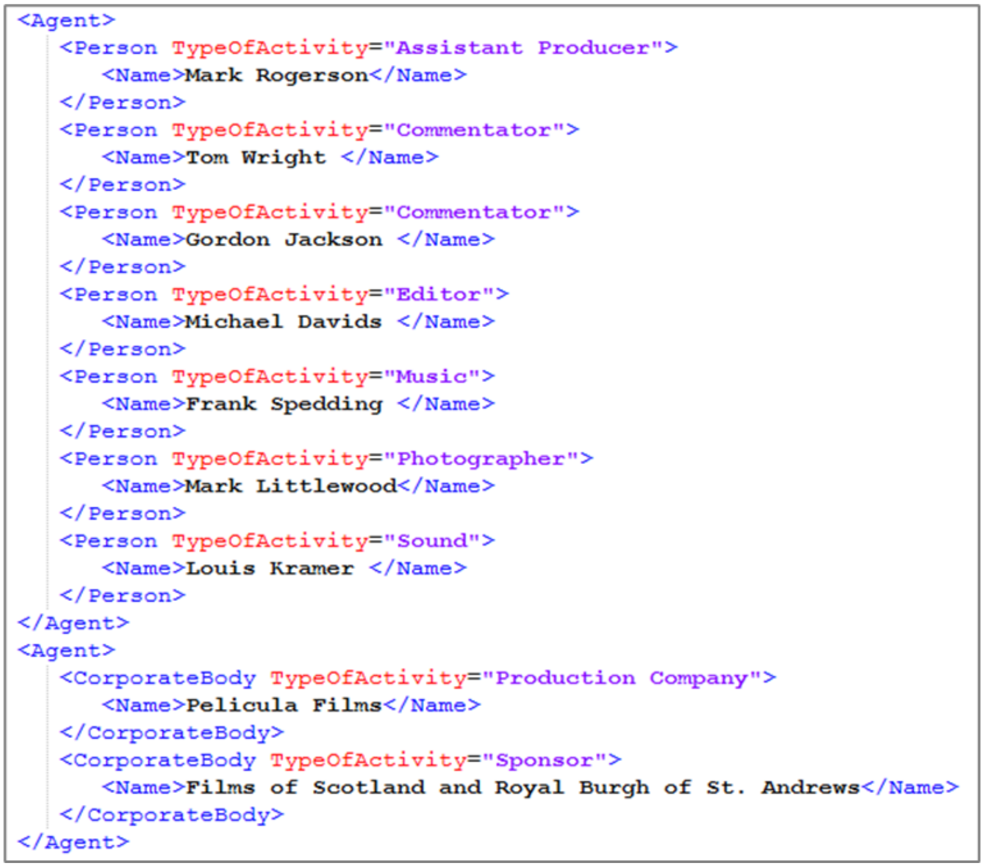
```
<descCredits>
    location sd. Louis Kramer
    production asst. Mark Rogerson
    ed. Michael Davids
    ph. Mark Littlewood
    asst. camera 'Kelly' Grey
    comm. w. Tom Wright
    comm. s. Gordon Jackson
    m. Frank Spedding
    cond. Marcus Dods
    Renaissance prints Orzel Studios
    m.p.s. St. Andrews University Choir conducted by Cedric Thorpe Davie
    m.p.s. St. Mungo's Choir conducted by Father Patrick Fitzpatrick
</descCredits>
```

Figure 3: Example of source MIA "Description Credits" field

[2] For more information on the National Bibliography of Scotland, see: Vincent, Helen; Cunnea; Paul; De Pretto, Alexandra (2018), *From Scottish Bibliographies Online to National Bibliography of Scotland: Reinventing a National Bibliography for the 21st Century*. Paper presented at IFLA WLIC 2018 in Session 244, available at: http://library.ifla.org/2275/ [accessed 16/03/2020]

[3] filmstandards.org, *Standards and Specifications*, available at: http://filmstandards.org/fsc/ [accessed 28/02/2020]

EFG uses a highly structured schema,[3] which requires the differentiation of agents as Person and agents as Corporate Body. In a similar fashion to the Aldis example, this data can be repurposed with the following result:



```
<Agent>
   <Person TypeOfActivity="Assistant Producer">
      <Name>Mark Rogerson</Name>
   </Person>
   <Person TypeOfActivity="Commentator">
      <Name>Tom Wright </Name>
   </Person>
   <Person TypeOfActivity="Commentator">
      <Name>Gordon Jackson </Name>
   </Person>
   <Person TypeOfActivity="Editor">
      <Name>Michael Davids </Name>
   </Person>
   <Person TypeOfActivity="Music">
      <Name>Frank Spedding </Name>
   </Person>
   <Person TypeOfActivity="Photographer">
      <Name>Mark Littlewood</Name>
   </Person>
   <Person TypeOfActivity="Sound">
      <Name>Louis Kramer </Name>
   </Person>
</Agent>
<Agent>
   <CorporateBody TypeOfActivity="Production Company">
      <Name>Pelicula Films</Name>
   </CorporateBody>
   <CorporateBody TypeOfActivity="Sponsor">
      <Name>Films of Scotland and Royal Burgh of St. Andrews</Name>
   </CorporateBody>
</Agent>
```

Figure 4: 'Magic' applied to the "Description Credits" field from example in fig. 3

This example demonstrates that following guidelines for content is not in itself enough: data that is hard to interpret for humans is practically unusable for machine processing, except in a limited capacity, such as relying on keyword searches for discovery. This transformation work is laborious and far from perfect: not everything has been included and not all issues are sorted. Further data analysis could remedy this problem to a certain extent, although a better scenario would be to alter the source data to reduce ambiguity and review how it is encoded. However, would that be enough?

Working with structured data

Manuscripts and Archives is another important collection within the National Library of Scotland. This dataset follows the Encoded Archival Description (EAD) standard, which is highly structured and appropriate to this type of collection and can be used effectively by a discovery system that has been designed for archival material. However, our primary discovery system cannot interpret EAD without significant intervention. Relevant EAD source elements need to be mapped to the most appropriate field of the schema used by our discovery system. In doing so, some elements don't transfer well – the hierarchies are partly lost, and with them the clear relationships between collections and sub-collection; in some cases the data is dropped completely.

This raises a number of questions: how much data is enough to ensure discoverability of the material we hold? Are we using the right standards and systems to describe and display our collections? Are we publishing our data in the right places?

From local to global

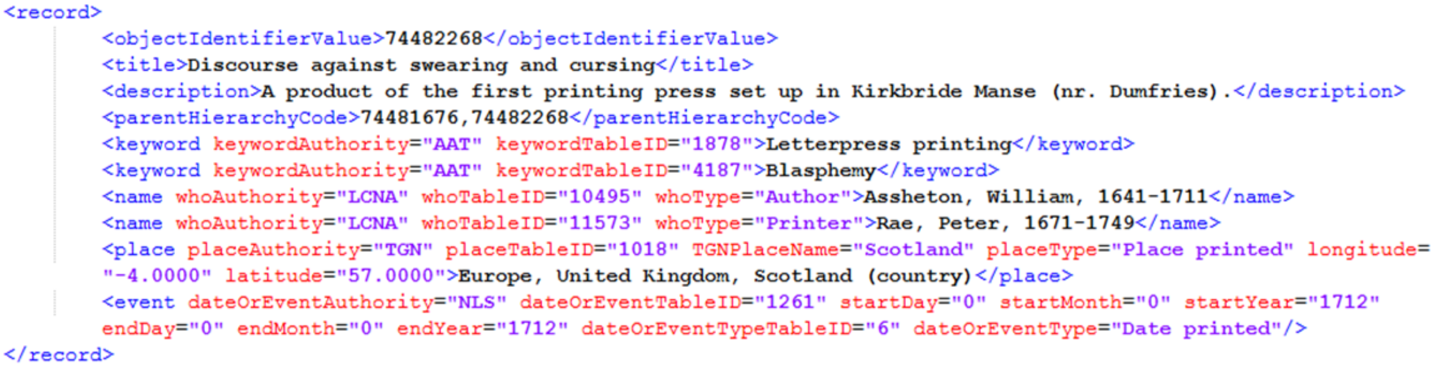
Over the years, information professionals have developed a diversity of metadata encoding and descriptive standards for specific purposes or types of material. Libraries are now faced with the increasingly pressing challenge of making their metadata available for web based services. There is no simple solution, but getting a clear understanding of some of these issues is a step in the right direction.

One example has been converting the Library's digitised collections metadata to the Europeana Data Model (EDM) in order to integrate these collections into the Europeana portal. The Library's Digital Objects Database (DOD) is an in-house system containing descriptive metadata for display and discovery of digital content on an interface exclusive to the National Library of Scotland. This database is constructed around the concept of records of a hierarchical nature.

EDM on the other hand, is an RDF based data model in which there is no concept of record or hierarchy, as it is built on triple statements connecting subjects to objects via properties. EDM is designed for the description of resources and their digital representations and adopts "an open, cross-domain Semantic Web-based framework that can accommodate the range and richness of particular community standards".[4] The model has the ability to express relationships between three core classes ("Provided Cultural Heritage Object", "Web Resource" and "Aggregation") and contextual classes (e.g. "Agent", "Place" or "Concept"). Unique Resource Identifiers (URIs) from controlled vocabularies are fully integrated in the model for properties and are strongly encouraged for as many other data elements as possible. Data in EDM can easily be processed by applications; the model is also flexible enough to integrate external sources of information.

[4] Europeana (2014), *Europeana Data Model primer*, p. 5, available at: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf [accessed 11/03/2020]

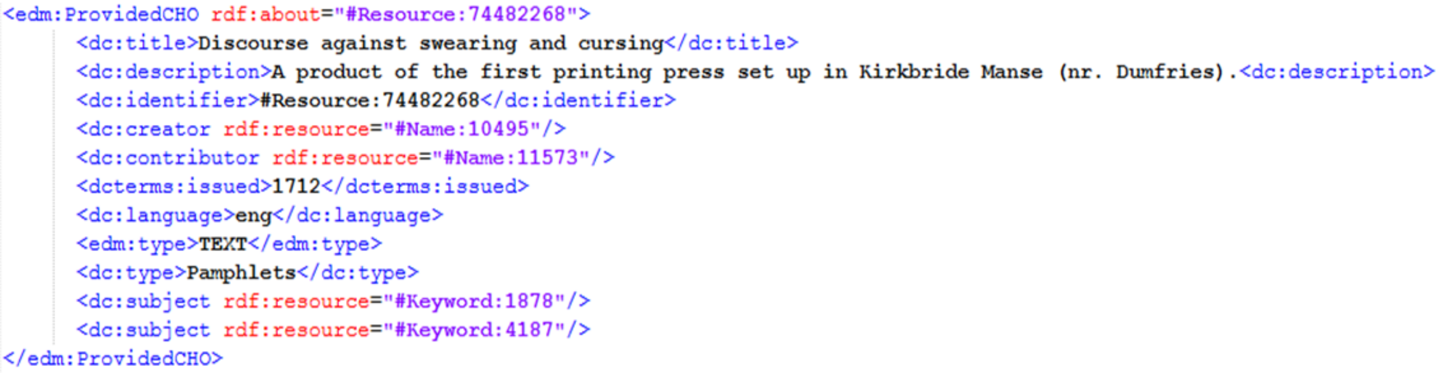
The lengthy but interesting exercise of mapping our DOD data to EDM brought into focus some of the limitations of working with a dataset based primarily on text and records. One of the challenges is to ensure data extracted from the DOD retains its meaning in a different context, as I explain with the example below:



```
<record>
        <objectIdentifierValue>74482268</objectIdentifierValue>
        <title>Discourse against swearing and cursing</title>
        <description>A product of the first printing press set up in Kirkbride Manse (nr. Dumfries).</description>
        <parentHierarchyCode>74481676,74482268</parentHierarchyCode>
        <keyword keywordAuthority="AAT" keywordTableID="1878">Letterpress printing</keyword>
        <keyword keywordAuthority="AAT" keywordTableID="4187">Blasphemy</keyword>
        <name whoAuthority="LCNA" whoTableID="10495" whoType="Author">Assheton, William, 1641-1711</name>
        <name whoAuthority="LCNA" whoTableID="11573" whoType="Printer">Rae, Peter, 1671-1749</name>
        <place placeAuthority="TGN" placeTableID="1018" TGNPlaceName="Scotland" placeType="Place printed" longitude=
        "-4.0000" latitude="57.0000">Europe, United Kingdom, Scotland (country)</place>
        <event dateOrEventAuthority="NLS" dateOrEventTableID="1261" startDay="0" startMonth="0" startYear="1712"
        endDay="0" endMonth="0" endYear="1712" dateOrEventTypeTableID="6" dateOrEventType="Date printed"/>
</record>
```

Figure 5: Original DOD data extract

Whilst we use controlled vocabularies for keyword, name and place information, as can be seen with the “keywordAuthority” attribute, this has been following traditional authority control management practices, which rely on using a unique form of name,[5] and in general on recording labels rather than identifiers; numbers referred to in this extract are DOD system numbers with no contextual information once extracted from the system. EDM uses established ontologies to define RDF properties (e.g. dc:creator using the Dublin Core schema) and encourages the use of Unique Resource Identifiers for classes. A compromise had to be found for our data, which was done by pre-fixing those with contextual information, e.g. #Resource:74482268, as demonstrated below:

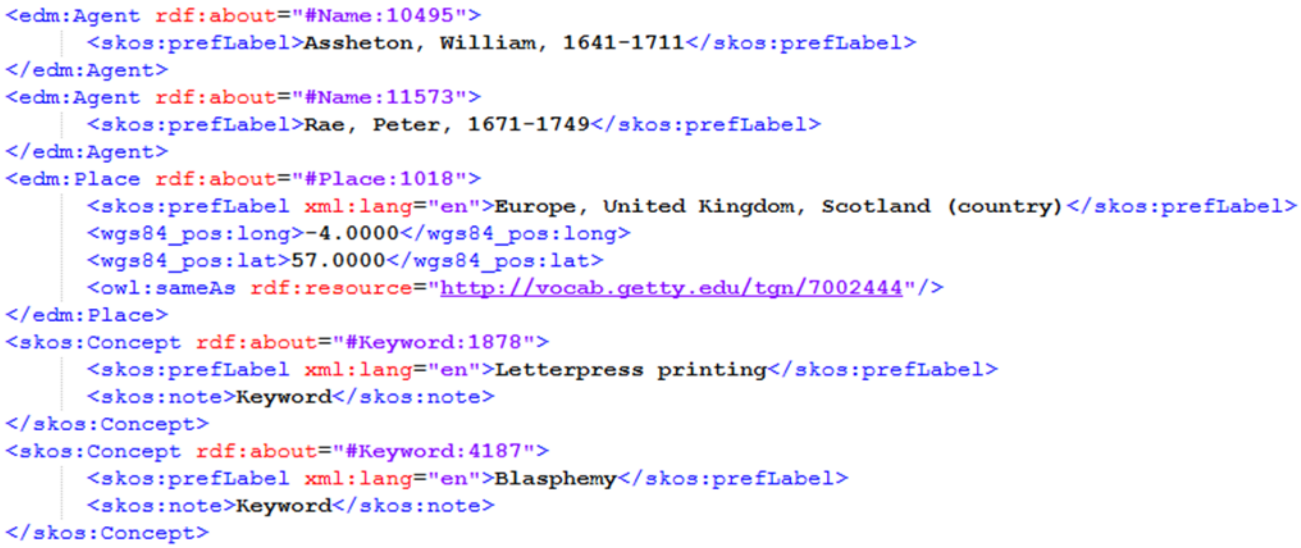


```
<edm:ProvidedCHO rdf:about="#Resource:74482268">
      <dc:title>Discourse against swearing and cursing</dc:title>
      <dc:description>A product of the first printing press set up in Kirkbride Manse (nr. Dumfries).<dc:description>
      <dc:identifier>#Resource:74482268</dc:identifier>
      <dc:creator rdf:resource="#Name:10495"/>
      <dc:contributor rdf:resource="#Name:11573"/>
      <dcterms:issued>1712</dcterms:issued>
      <dc:language>eng</dc:language>
      <edm:type>TEXT</edm:type>
      <dc:type>Pamphlets</dc:type>
      <dc:subject rdf:resource="#Keyword:1878"/>
      <dc:subject rdf:resource="#Keyword:4187"/>
</edm:ProvidedCHO>
```

Figure 6: Extract of EDM converted DOD data for the ProvidedCHO class

[5] Armitage A., Cuneo M.J., Quintana I. & al. (2020), *ISNI and traditional authority work*, in: Italian Journal of Library, Archives and Information Science, 11, 1, 2020, p. 154, available at: https://doi.org/10.4403/jlis.it-12554 [accessed 16/03/2020]

The data above describes the "Provided Cultural Heritage Object" core class, which includes references to contextual classes such as Keyword, Agent and Place, which are represented as follows:



```
<edm:Agent rdf:about="#Name:10495">
      <skos:prefLabel>Assheton, William, 1641-1711</skos:prefLabel>
</edm:Agent>
<edm:Agent rdf:about="#Name:11573">
      <skos:prefLabel>Rae, Peter, 1671-1749</skos:prefLabel>
</edm:Agent>
<edm:Place rdf:about="#Place:1018">
      <skos:prefLabel xml:lang="en">Europe, United Kingdom, Scotland (country)</skos:prefLabel>
      <wgs84_pos:long>-4.0000</wgs84_pos:long>
      <wgs84_pos:lat>57.0000</wgs84_pos:lat>
      <owl:sameAs rdf:resource="http://vocab.getty.edu/tgn/7002444"/>
</edm:Place>
<skos:Concept rdf:about="#Keyword:1878">
      <skos:prefLabel xml:lang="en">Letterpress printing</skos:prefLabel>
      <skos:note>Keyword</skos:note>
</skos:Concept>
<skos:Concept rdf:about="#Keyword:4187">
      <skos:prefLabel xml:lang="en">Blasphemy</skos:prefLabel>
      <skos:note>Keyword</skos:note>
</skos:Concept>
```

Figure 7: Extract of EDM converted DOD data for Agent, Place and Concept

In an ideal world, rather than using local identifiers, each contextual class would link to the external URI available from AAT, LCNA and TGN vocabularies, which we use in our source database. Such identifiers have not been routinely recorded in our systems in the past, but we now record TGN identifiers for place information and this can be seen in the owl:sameAs property.

Making use of controlled vocabularies, therefore, whilst important, is not enough, given the increased need for connectivity and efficiencies. Seeing our digitised collections data in EDM demonstrates the potential of semantic web enrichment and linking resources internally and externally in a way not possible with our current set-up. Collections become much more accessible to web-based search engines and services, with the added benefit that machines as well as humans can interpret this data.

As highlighted by one of the OCLC Project Passage's participants, "*Querying existing bibliographic and authority data has always helped catalogers understand 'what is already out there' and subsequently, add to it. We will just be doing this task more efficiently by a singular factual statement on entities/relationships rather than by endlessly populating strings over and over on a record-by-record basis.*"[6]

[6] Knudson Davis Kalan (2020), A*n insider's look at "Project Passage" in seven linked data lessons, six constants, five changes … and four webcomics*, available at: http://www.oclc.org/blog/main/insiders-look-at-project-passage/?utm_source=SFMC&utm_medium=email&utm_content=vol23-no5-feature-article-join-oclc-at-pla-2020&utm_campaign=oclc-abstracts-vol23&utm_term=OCLC%20Abstracts_COMM [accessed 10/03/20]

Concluding thoughts

Libraries are rich in descriptive metadata, which is, after all, the means by which a library delivers a wide range of services. There is, however, a much greater potential. Libraries and related disciplines have been working with structured data using a variety of standards for a long time, but if, as metadata librarians, we don't embrace new developments and technologies, we replicate the same problems: isolation, miscommunication (ambiguity) and lack of efficiency: we are essentially lost in translation.

In its new metadata strategy, the British Library highlights the fact that legacy infrastructure and standards have to some extent become hindrances to the opportunities presented by metadata. As efficiencies become increasingly dependent on automated workflows, their aim is for a more integrated approach based around a uniform, system-neutral metadata model for all collections. This is also motivated by the opportunity for a "coherent resource discovery" user experience.

If you have ever spent a number of years learning a foreign language at school, you will no doubt remember the requirement to learn the rules of grammar methodically, which you then applied more rigidly than the native speaker. Then comes the day when you are comfortable enough in your knowledge of the language that you can let go and, without fully compromising the rules, be less constrained and converse more fluidly.

Our discovery systems may provide an adequate service, but they still rely on outdated library based encoding and formats that require significant human intervention for customising, transforming, linking or maintaining. Being more resilient for the future should start by reviewing existing systems and practices, coordinating which standards are used, and reviewing data models. This is something that the National Library of Scotland has started doing through developing metadata principles and a standards matrix, and setting up a working group to review use of unique identifiers for specific entities. More remains to be done to expand our horizons, not least develop our own metadata strategy.

Let's not lose sight that, for libraries, metadata is the foundation for effective communication about amazing resources.

[7] The British Library, *Foundations for the future: The British Library's Collection Metadata Strategy 2019-2023*, p.3, available at: https://www.bl.uk/collection-metadata/strategy-and-standards [accessed 06/03/20]

Library data as linked open data

Jason Evans, National Wikimedian at the National Library of Wales

Libraries have changed. Around the turn of the last century new public libraries were being built in towns and cities all over Britain. The Scottish-American businessman and philanthropist Andrew Carnegie funded more than 2,500 new libraries around the world, and here in Wales the National Library was established in 1907. For many years the library catalogue was no more than ink on paper: card indexes in deep wooden drawers, painstakingly compiled by humans. Both the appetite for building libraries and the way in which they operate have changed somewhat in the intervening century but the core purpose of the library is very much the same – to give access to knowledge, literature and learning.

At the National Library of Wales we still use some of our old card indexes. Archive collections tend to be indexed by category and subcategory and many have been annotated by catalogers, noting related documents. In its simplest form, this is linked data – the ability to make connections across collections based on a number of different common factors such as author, subject or place.

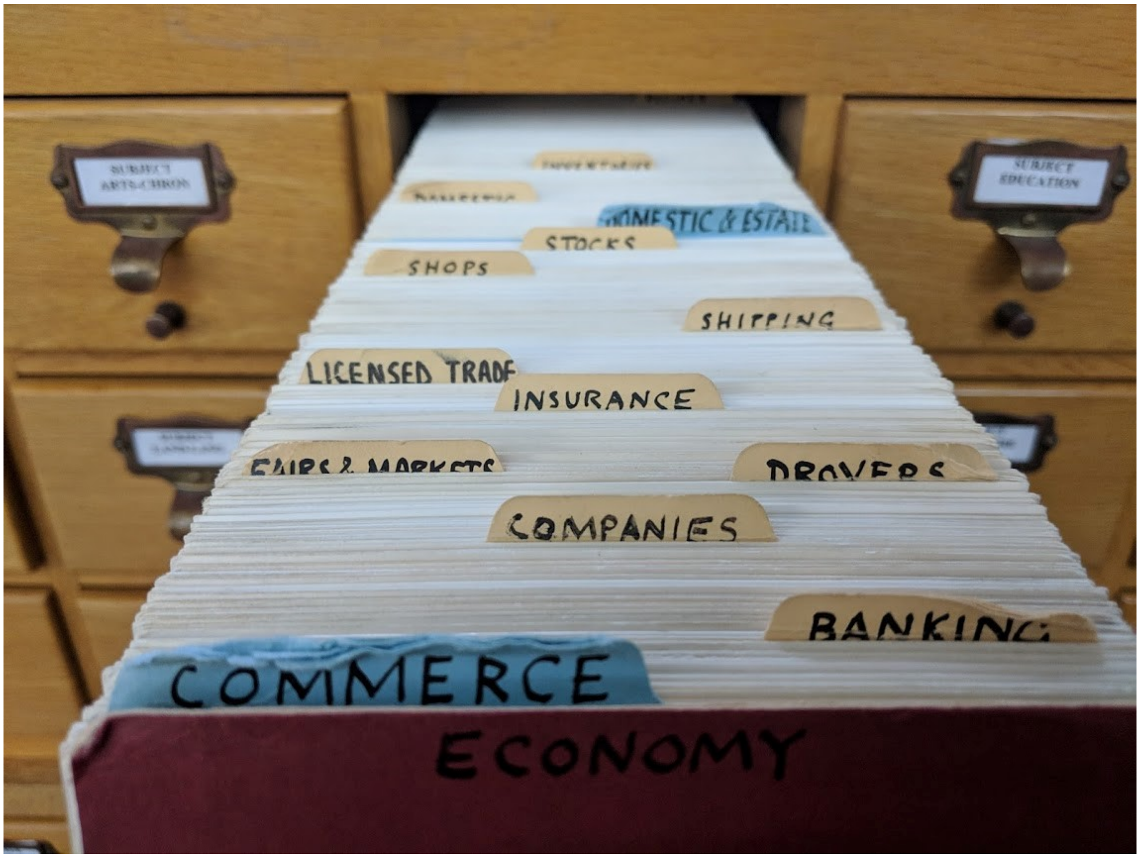
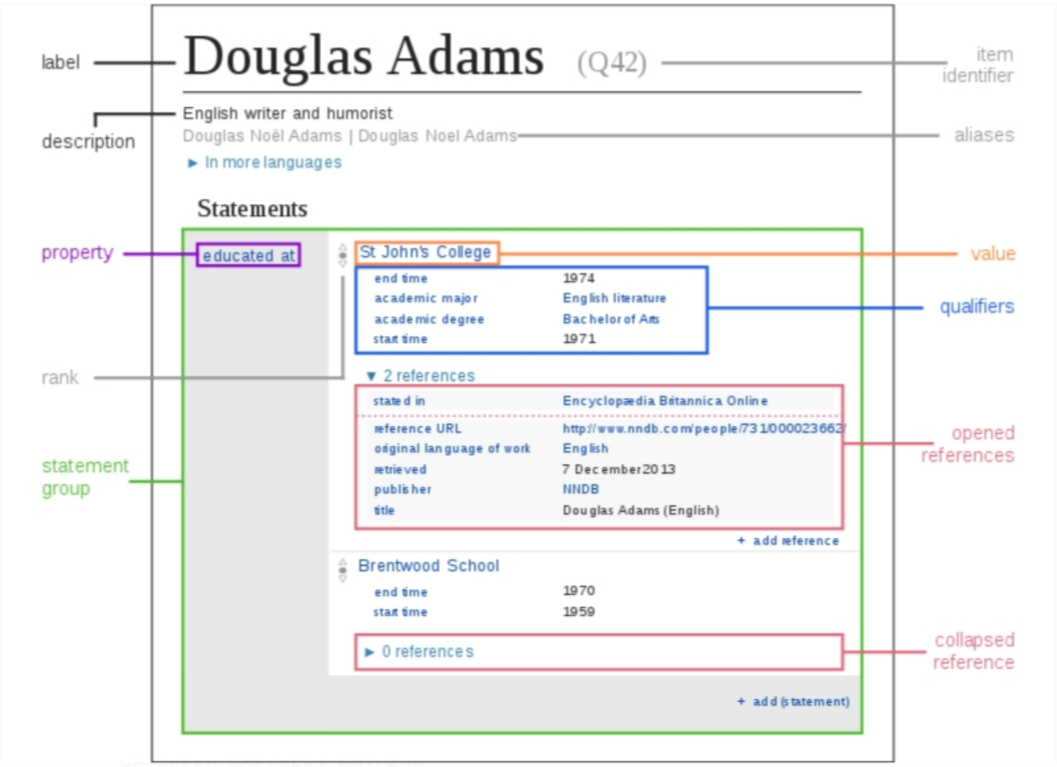


Figure 1: An old card index at the National Library of Wales

So, if our aim is to give the best access possible to our users, both our MARC21 data for printed material and our Dublin Core archive and manuscript data have certain shortcomings. This is both down to the way the data is structured and the way we give access to it. At the National Library of Wales we have been exploring the benefits and challenges of converting to linked open data. To date this has been done by experimenting on a reasonably small scale using the established infrastructure of Wikidata.

Wikidata is part of the Wikimedia Foundation's family of websites – you may have heard of its flagship site Wikipedia. Much like Wikipedia, Wikidata is open, editable, and reusable by anyone. It is part of the Wikimedia Foundation's vision of a world in which everyone has access to the sum of all human knowledge, and, as the name suggests, Wikidata's contribution is in the form of data. Originally created about 6 years ago to help connect related Wikipedia articles in different languages, Wikidata has flourished into a massive open data set with data about just about everything, from the human genome to people, places, events, literature, and all sorts of other 'things'. At the time of writing, Wikidata describes nearly 85 million 'things'. What makes Wikidata so powerful, other than the fact that institutions and individuals can contribute or make use of the data freely, is that it is linked data, rather than text based data – it's 'things' not 'strings'. Behind the interface, Wikidata looks very much like RDF.[1] Each item has a unique identifier (a Q number) along with a label and description which can be added in over 300 languages, making this a multilingual dataset. Items are then described using an unlimited number of statements, also known as triples. The dataset already contains over a billion such statements.

Image from SPARQL/WIKIDATA Qualifiers, References and Ranks



Douglas Adams (Q42)

English writer and humorist

Douglas Noël Adams | Douglas Noel Adams

► In more languages

Statements

educated at

St John's College

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971

▼ 2 references

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)

+ add reference

Brentwood School

end time	1970
start time	1959

► 0 references

+ add (statement)

```
?country wdt:P463 wd:Q458.
SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
```

Figure 2: Annotated example of a Wikidata Item

[1] https://www.wikidata.org/wiki/Wikidata:Relation_between_properties_in_RDF_and_in_Wikidata

By converting catalogue data to linked open data on Wikidata we are essentially taking the metadata 'fields' and matching them to Wikidata 'properties'; in other words we are the converting text 'strings' that inhabit those metadata fields and matching them to Wikidata items (things). For example, if the value of a metadata field for place of publication is 'Aberystwyth' we will match this to the Wikidata item about Aberystwyth (Q213154). This means that our data no longer simply states 'Aberystwyth' as the place of publication in a human readable way, but creates a connection, in a machine-readable way, to all the data Wikidata holds about Aberystwyth, such as the type of town, its date of inception, its population, and its coordinate location. Wikidata also acts as a hub for external identifiers, and the item for Aberystwyth includes information such as the WorldCat ID, VIAF ID, and the Library of Congress authority ID for the town.[2]

As you might imagine, transforming standard metadata effectively into this format has its challenges. Mapping metadata field types to the relevant Wikidata properties is pretty straight-forward, but matching text strings to the correct Wikidata items takes more time. Wikidata has 13 items called 'Aberystwyth'. Most are paintings, but there is also an item for a hymn and one for a 19th century merchant ship. There are tools which allow you to filter out these irrelevant items, such as OpenRefine, but often an element of manual checking is required. This is particularly true when dealing with creators of works – untangling 50 different John Joneses and ensuring each has their own data item with at least some biographical data is no small task, trust me. Often this problem is compounded by a lack of consistency in the way data is presented in MARC fields.

Despite the challenges we have seen some interesting results. We took data for the Peniarth Collection – some 540 medieval manuscripts – and converted it to linked data. Some data fields mapped easily to the new format, but much of the useful descriptive data for the items formed a text description in a 'scope and context' field. This was teased apart using patterns in the text to identify scribes, authors, and works contained in the manuscripts. We found that many of the authors and scribes were already in Wikidata thanks in part to our previous work to share data for the Dictionary of Welsh Biography. Once all this data was matched up to Wikidata, we found that we had access to a much richer dataset than before. Using Wikidata's query service you can query every element of the data. So, for example, we could easily isolate works by an individual scribe, or discover which scribes worked on the same manuscripts. We could generate lists of manuscripts by genre and main subject, as well as access newly connected information such as biographical data about authors. So, for example, we could return a list of items in the collection authored by women born in the 15th century, or manuscripts that included works about the Roman Empire written in Welsh. And because Wikidata allows institutions to add their own identifiers to items on Wikidata we discovered that other institutions held different versions of some of the manuscripts in our collection, or other manuscripts with the same authors or scribe. The opportunity for research is massively increased both by the nature of the structured data and by the ease of access to the data through Wikidata's query service and suite of visualisation tools.[3]

[2] https://www.wikidata.org/wiki/Q213154

[3] https://blog.library.wales/treasured-manuscript-collection-gets-the-wikidata-treatment/

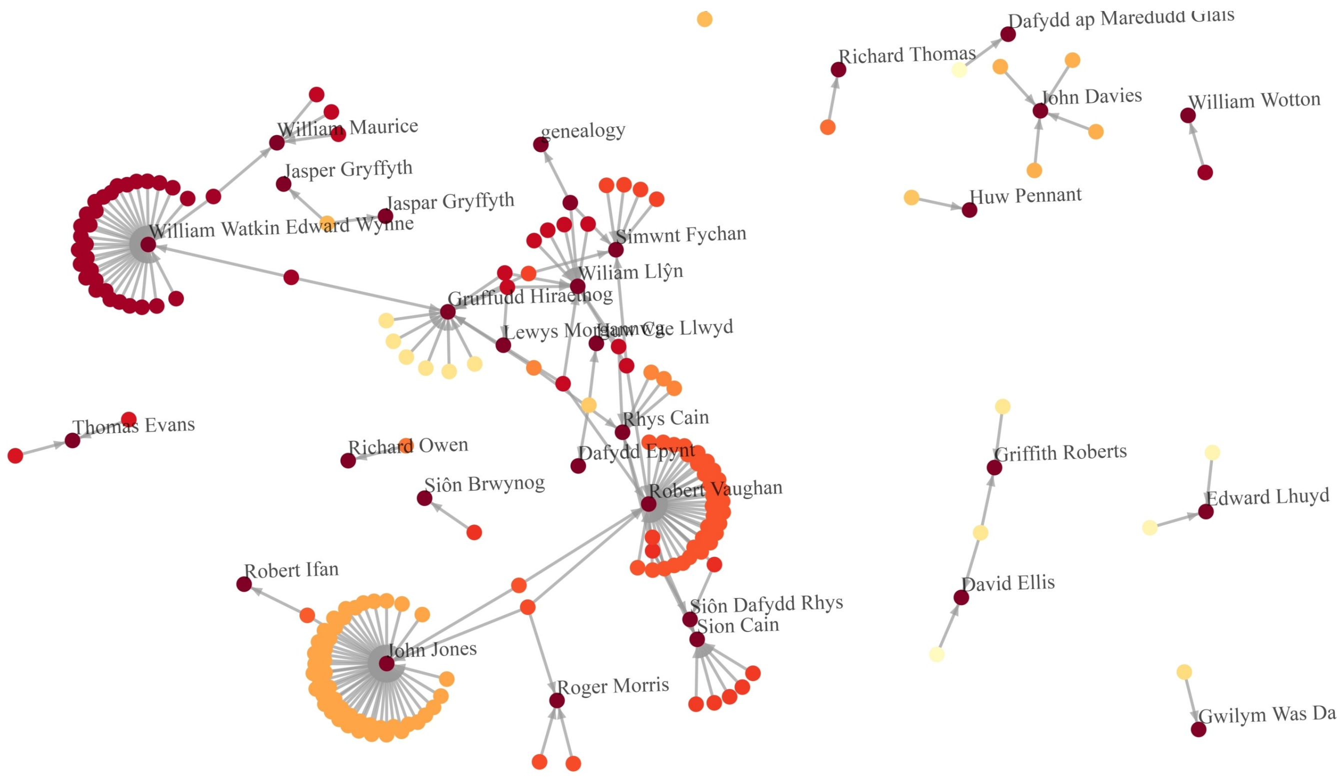


Figure 3: Scribes of Peniarth Manuscripts and the works with which they are associated

As well as sharing data through Wikimedia platforms we have also contributed to Wikipedia and Wikimedia Commons, where we have shared around 20,000 openly licenced images of photographs, artworks and other digitised content. Wikidata can then be used to describe this content using linked data. Again, the process of transforming our data led to an enrichment of our data record. Matching tags of places depicted in images to place items on Wikidata meant we gained access to coordinate data, and so we could explore the collection on a map for the first time. We also gained access to a wealth of information about the artists, the types of things depicted in images, and other institutions who also held either works depicting those things, or works by the same artists.

We have also seen other positive outcomes from making our data open. Wikidata applies a CC0 licence to all its data, and most of the images we have shared to Wikimedia Commons are in the public domain. This means that our content can be freely reused by anyone. We have seen our content used in thousands of Wikipedia articles, garnering millions of views every month. Authors, historians, curators, and the media have used our free open content to enrich their own resources, but increasingly we are seeing that by ‘round tripping’ our improved, structured open data we can develop and improve our own online services.

We worked with volunteers to develop a prototype for a linked data portal for exploring our open digital images, by adapting existing open source software. The result is the ‘Dwynwen’ website which allows users to search over 17,000 images in ways that are simply not possible with our standard catalogue data. You can even view content on a map based on places depicted in the images.

We will also shortly be releasing an interactive timeline for the Dictionary of Welsh Biography which is powered by Wikidata, open images from Wikimedia Commons, and text from Wikipedia.

As Wikidata has grown it has increasingly become a hub for describing publications. Scholarly articles in particular now make up 31% of Wikidata.[4] One of the driving forces behind the use of Wikidata in this way is to improve accuracy of content on Wikipedia. Creating a central, open, structured dataset of scholarly content makes it easier for Wikipedia editors to locate and cite reliable content for Wikipedia. Discussions as part of the WikiCite initiative suggest that in the long term all citations on Wikipedia could be standardized across all languages using linked data. It would mean Wikipedia users could access all those external identifiers used on Wikidata to see who holds copies of a book or paper, where it was available online, and how it was licenced for reuse. With this in mind the National Library of Wales recently shared over 30,000 catalogue entries from the Welsh bibliography on Wikidata.

This task presented a number of challenges. Because each piece of data has to link to another Wikidata item, text strings for publishers and authors either had to be matched to existing data items or the data had to be added. We found that one publisher might appear in dozens of different formats in our metadata. Some records gave publisher names in Welsh, others in English. This meant a lot of checks and processing using OpenRefine and Excel in order to clean the data. Hundreds of Welsh publishers and authors were added to Wikidata, and whilst finding, disambiguating, and locating information about all the publishers was a manageable task, the same cannot be said for authors. 13,000 works are now linked to items for over 2000 unique authors. However, we were not able to identify and match the remaining authors in the time we had.

Many of the authors were identified by matching our data with other datasets. For example, 4536 items were matched to OCLC bibliographic records,[5] 11386 have universal ISBN numbers,[6] and 6941 were matched to Open Library records.[7] This approach of connecting with different datasets is key to cleaning up and converting to linked data. Restructuring our data in this way at scale will depend on close collaboration across the sector.

Although the transition to linked data in libraries is still a fairly new idea, there are some notable examples of the use of not only linked data, but also open and collaborative data, being used to enrich user experiences. The Biblioteca Nacional de España (BNE) in Spain has created linked data for its entire catalogue, allowing them to create a powerful search engine which incorporates open text from Wikipedia.[8] At the National Library of the Netherlands, the OCR text of their online newspaper archive has been enriched with Wikidata entities, improving search and linking users to a wealth of additional information about people, places and events identified within the text. The National Library of Sweden has also begun sharing their national bibliography with Wikidata.

There seems to be a growing acceptance that linked data will form part of the next chapter in the ever changing landscape of library and information services. It offers more powerful search and discovery tools, more consistent and complete data, and better access for users. It will also allow us to begin connecting and aligning data between institutions, creating one huge global dataset. The open and collaborative elements of a system like Wikidata mean that we can work together with our partners and with our communities to develop, enrich, and research our data in new ways. This collaborative approach also helps mitigate some of the obvious and inevitable challenges the adoption of linked data presents, and allows us, as a sector, to pool resources in order to improve data, and ultimately to continue to serve as essential providers of knowledge.

[4] https://www.wikidata.org/wiki/Wikidata:Statistics

[5] Wikidata query for NLW works with OCLC ids

[6] Wikidata query for NLW works with ISBN numbers

[7] Wikidata query for NLW works with Open Library ids

[8] The BNE 'Datos' linked data portal

[9] van Veen T., Lonij J., Faber W.J. (2016) Linking Named Entities in Dutch Historical Newspapers. In: Garoufallou E., Subirats Coll I., Stellato A., Greenberg J. (eds) Metadata and Semantics Research. MTSR 2016. Communications in Computer and Information Science, vol 672. Springer, Cham

MARC transformed: MARC and XML – the perfect partnership?

Heather Rosie, EThOS Repository Metadata Manager, The British Library

Introduction

I first met a very British version of MARC (Machine Readable Cataloguing) in 1983, straight out of university. I didn't know anything about cataloguing, indexing, classification, or data. MARC made sense of it all. AACR2 (Anglo-American Cataloguing Rules, 2nd Edition) was impenetrable without MARC as a framework. LCSH (Library of Congress Subject Headings) and DDC (Dewey Decimal Classification) seemed like a foreign language. MARC gave the rules structure and validity. MARC made the rules live. And, as time progressed, MARC matured and developed into MARC 21. For me, this was the beginning of a life-long love affair with MARC.

But, in 1996, MARC's world was turned upside down when XML (Extensible Markup Language) appeared. MARC was no longer the only kid on the block. And XML knew all about the Internet, linked data, and OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). XML was in the cloud; MARC was firmly on the ground. Users were falling out of love with MARC. It was time to re-evaluate the qualities that made MARC great in the first place – and some help was needed. Luckily, help was at hand in the form of two very clever desktop software packages, MarcEdit and MARC Report. These invaluable tools have made MARC pivotal in the world of bibliographic data transformation. Without MARC, EThOS (E-Theses Online Service) would not be the unique aggregation of thesis data that it is today. EThOS makes it easy to Find, Identify, Select and Obtain[1] a UK doctoral thesis because it is based on good quality, de-duplicated bibliographic data. And this data originated as XML, was transformed by MARC, and then returned to XML. The perfect partnership.

MARC, XML and Dublin Core

So how do we get MARC to play in an XML world? The solution is not to convert MARC ISO 2709 to MARC XML; if consumers of MARC XML are not familiar with the MARC 21 Bibliographic Format or cataloguing standards (AACR2; RDA – Resource Description and Access) the data won't be any more re-usable than the original MARC format. But the answer is to convert MARC to XML – just not 'MARC XML'. This is where Dublin Core has a role. Dublin Core does not have the complexities of MARC 21, but its simplicity has made it the de facto format for data exchange in the Institutional Repository landscape. Unfortunately, this usually means the ‘simple' Dublin Core format, and this *is* too restrictive. It is possible, however, to develop a more sophisticated format using all the elements available in Qualified Dublin Core (QDC), and adding any further required elements that are not defined in the QDC schema. This is the solution developed by *EThOS*, the British Library's national e-thesis service.

[1] https://www.oclc.org/research/activities/frbr.html

EThOS (Electronic Theses Online Service)

EThOS was launched in 2009 to replace a document supply service based on microfilm reproductions. The initial aim was to digitise the hard copy collections of each UK university. Demand for digital material was many times higher than the old microfilm service and, at its inception, EThOS focused on digitisation, somewhat to the detriment of resource discovery. The indexing and discovery side of the service was based on the migration of around 250,000 MARC 21 catalogue records from the British Library's Integrated Library System, Aleph, via a newly developed Dublin Core-based XML schema – UKETD_DC (UK E-Theses and Dissertations Dublin Core). The EThOS Application Profile for this schema uses thirteen QDC elements (some with attributes to make them more specific) and nine newly defined ‘uketdterms' elements particular to the ‘thesis' material type (e.g. Qualification Name and Qualification Level). The schema could accommodate all the MARC elements in the source data (apart from coded fields) and included elements not present in catalogue records (e.g. funding bodies).

The new schema was developed to facilitate the harvesting of metadata from every UK university, because the ultimate aim of EThOS is to index and provide free, immediate access to the full-text digital content of every doctoral thesis awarded in the UK. At the time of writing (May 2020) EThOS holds 555,000 records and provides access to 342,000 digital theses (60%).

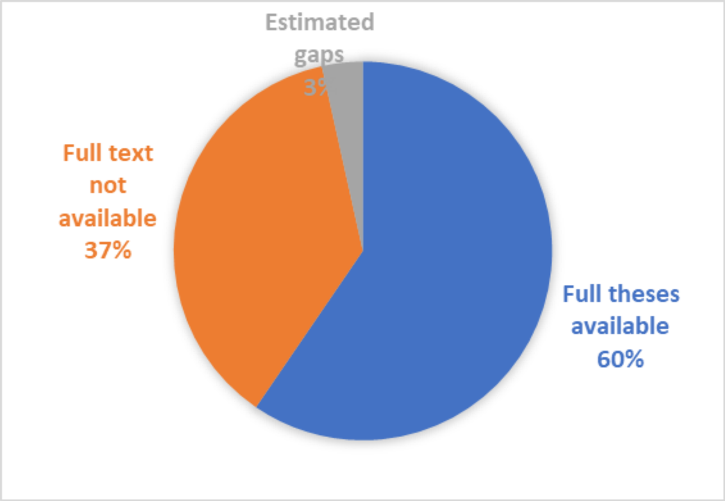


Figure 1: Proportion of UK doctoral theses available in digital form (May 2020)

OAI-PMH and MarcEdit

The established standard for metadata exchange in the Open Access community is OAI-PMH. This is a web-based XML transfer protocol. It involves constructing an HTTP (Hypertext Transfer Protocol) request in a web browser to generate an XML response that can be downloaded, transformed, and edited. For those in the library community more familiar with MARC than XML, the obvious data transformation is XML to MARC. And there is a fabulous tool available for just this job – MarcEdit.

MarcEdit is a 'free to download' software application for editing and processing MARC data. It is in use by thousands of people worldwide and includes a host of utilities for bulk data processing. One of these is an OAI Harvester. OAI-PMH is intended to be a low-barrier protocol for data exchange, which essentially means specialist software is not required. It is possible to harvest metadata using just a web browser, but constructing the syntax can be complicated, and the data is returned in 'chunks' (i.e. separate files of less than 1,000 records) due to flow control measures used to "throttle incoming requests from harvesters".[2] The data is also returned as XML and, if you want to use MARC processing utilities, you need that data transformed to MARC.

The MarcEdit OAI Harvester makes the harvesting and transformation of XML data very easy. It includes the following elements:

1. A template to complete all the details required to construct a harvesting query.
2. Default XSLT (eXtensible Stylesheet Language Transformations) to transform the harvested XML data to MARC 'on the fly'. The list of style sheets can be modified and extended according to the user's requirements.
3. A behind-the-scenes utility to concatenate the 'chunks' of harvested data into a single MARC file.

Running a harvest takes seconds – and the result is a single MARC file that cataloguers and bibliographic metadata specialists can review and edit for their own specific purposes. Terry Reese, the developer of MarcEdit, refers to MARC in the context of OAI harvesting as sitting at the centre of a wheel in a 'spoke-and-wheel design'. In his article on 'automated metadata harvesting' he says "… This architecture allows metadata conversions to be created without the need to know directly how the individual metadata elements relate to elements within different schemas. Once a new spoke has been added to the wheel, it becomes crosswalk-able to any other spoke on that wheel."[3] In other words, MARC is the transitionary format that sits between one XML schema and any other XML schema.

Quality Control, Bulk Data Editing and MARC Report

Once the transformation from XML to MARC has been done, the resulting metadata file needs to be reviewed and edited to meet the requirements of the product or service it is to be used for, and a great tool for doing just that is MARC Report.

MARC Report is a 'paid for' desktop application originally developed for the quality control of MARC 21 bibliographic records – hence the name (it was an error 'reporting' system). It was developed by Deborah and Richard Fritz of TMQ (The MARC of Quality), who are also the authors of *MARC21 for Everyone: A Practical Guide*.[4] The software contains thousands of automated checks on the structure and content of MARC 21 bibliographic records created using the AACR2 or RDA cataloguing rules. It is also highly configurable by the end user.

---

[2] https://www.openarchives.org/OAI/2.0/guidelines-repository.htm (2.5 Flow Control)

[3] Reese, T. (2008). "Automated Metadata Harvesting: Low-Barrier MARC Record Generation from OAI-PMH Repository Stores Using MarcEdit". *Library Resources & Technical Services*, 53(2), pp.121-134. ISSN (0024-2527). DOI 10.5860/lrts.53n2.121

[4] Fritz, Deborah A. (2002). *MARC21 for Everyone: A Practical Guide*. ALA Editions

In addition to quality checks on individual bibliographic records, the software contains multiple utilities for bulk data processing, including MARC 21 to XML and XML to MARC 21 transformations using XSLT. The utilities enable a user to:

- Split and concatenate MARC files
- Import and export text files
- Import Excel files
- Sort and de-duplicate MARC records
- Generate statistical reports on the content of a MARC file (number of occurrences of tags, length of fields, etc.)

Two of the main features of MARC Report are 'MARC Review' and 'MARC Global' (the latter is an 'add-on' piece of software). MARC Review is a TDM (Text & Data Mining) tool for MARC files, particularly useful when the files include abstracts, as is the case with EThOS data. Any file of MARC records, no matter the size, can be analysed using 'patterns' based on the content of selected MARC tags. Analysis is by text string (rather than key word) and includes the ability to incorporate Perl Compatible Regular Expressions.

MARC Global is a bulk data editing tool. Fields and subfields can be added, deleted, copied, changed, and re-ordered. Data within tags can be changed, normalised, and de-duplicated, again with the option to use regular expressions. Automated 'scripts' for batch processing can be written without the need for programming knowledge.

The combination of individual record review and edit, bulk data analysis and manipulation, import and export of different data formats, and user configurability makes MARC Report/MARC Global a very powerful MARC processing software suite.

Repository Metadata vs. Catalogue Records

Data transformations from XML to MARC to XML are integral to any transfer of data between different institutional systems. For universities and national libraries, or other related research institutions that have both a catalogue (based on MARC records) and an institutional repository (based on DC XML records), the ability to transfer records between systems is a key requirement. Data is not stored in either of these systems in the same format it is exchanged; the format for data storage is determined by system developers. The expertise for the export, sharing, and import of data, therefore, lies with metadata experts, not software developers.

The British Library has always included bibliographic MARC records in its catalogue.[5] Implementing a specialist service for UK PhD theses did not mean that records would no longer be included in this catalogue, but the primary source of data for the discovery and retrieval of doctoral theses changed from a MARC-based system to a DC-based system. The main mechanism for populating the thesis database changed from individual record creation by cataloguers to bulk record load from institutional repository harvests. The data flow for EThOS is now:

OAI UKETD_DC XML => MARC21 ISO2709 => UKETD_DC XML => MARC21 ISO2709 (RDA)

The systems exporting, transforming, and importing this data are:

Institutional Repository => MarcEdit => MARC Report/MARC Global => EThOS and Aleph/Primo

[5] Explore the British Library: http://explore.bl.uk/primo_library/libweb/action/search.do?vid=BLVU1

This type of complex workflow is only possible because it incorporates quality review and editing as part of the transformation. The data requirements for each of these systems and services are different, even though there are common elements. The requirements of an individual university repository differ from the requirements of an aggregating service like EThOS which, in turn, differ from the requirements of a catalogue. The policies, standards, and rules that determine the requirements are also very different, especially for catalogue records where the standards are more exacting. The container (MARC or XML) is actually a pretty insignificant aspect of the data transformation – it is merely the vehicle that allows the data to be accepted by the system it is being transferred to or from. It is a common misconception that data transfers only require a change of 'container'. Whilst the data may be transferred, the end result – the service or product the data underpins – will be severely compromised if the data being transferred is not reviewed and edited to meet the requirements of that service or product.

Conclusions

How do we evaluate the success of a web-based service that has largely been built on data transformations between MARC and XML? Number of visits to the website? Number of downloads? Positive feedback? All of these? If it is the latter, then EThOS ticks all the boxes. But how about novel use of the aggregated metadata set for research purposes? In her 2016 article,[6] Sara Gould, British Library Repository Services Lead, mentions three case studies that used the EThOS metadata set to generate new knowledge and data (undertaken by the Alzheimer's Society, Royal Society of Chemistry and the FLAX Language Learning Project). More recently, a collaboration between EThOS, the Library's Technical Lead for the UK Web Archive, and Dr Peter Murray-Rust (chemist and open access advocate) has looked at the value of bringing metadata and full-text together to surface new information about coronavirus and the current Covid-19 pandemic.[7] This type of research only generates valid and verifiable results if the (meta)dataset is accurate, consistent, and comprehensive as far as is possible. EThOS is pretty unique in the arena of aggregating services with regard to the 'cleanness' of the dataset, as noted by Markus Hauru in a 2020 GitHub post ("The first observation to make is that the data is remarkably clean").[8] Other aggregating services, such as NDLTD, CORE, and even Google, focus on volume of data rather than integrity of the data corpus. In the case of EThOS, using MARC as a transitionary format has produced a dataset capable of playing a small, but significant, role in a world of massive aggregators – a fitting tribute to a format that is still proving its value today, almost 60 years after it was first invented.

[6] Gould, S. (2016). "UK theses and the British Library EThOS service: from supply on demand to repository linking", Interlending & Document Supply, 44(1), pp. 7-13. ISSN (0264-1615). DOI 10.1108/ILDS-10-2015-0033

[7] Jackson, A. (2020). “Bringing Metadata and Full-text Together”. British Library Digital Scholarship blog. https://blogs.bl.uk/digital-scholarship/2020/05/bringing-metadata-full-text-together.html

[8] Hauru, M. “Exploring trends in UK academia using PhD thesis metadata”, https://github.com/mhauru/EThOS-analysis/blob/master/analysis.ipynb (accessed 25 May 2020)

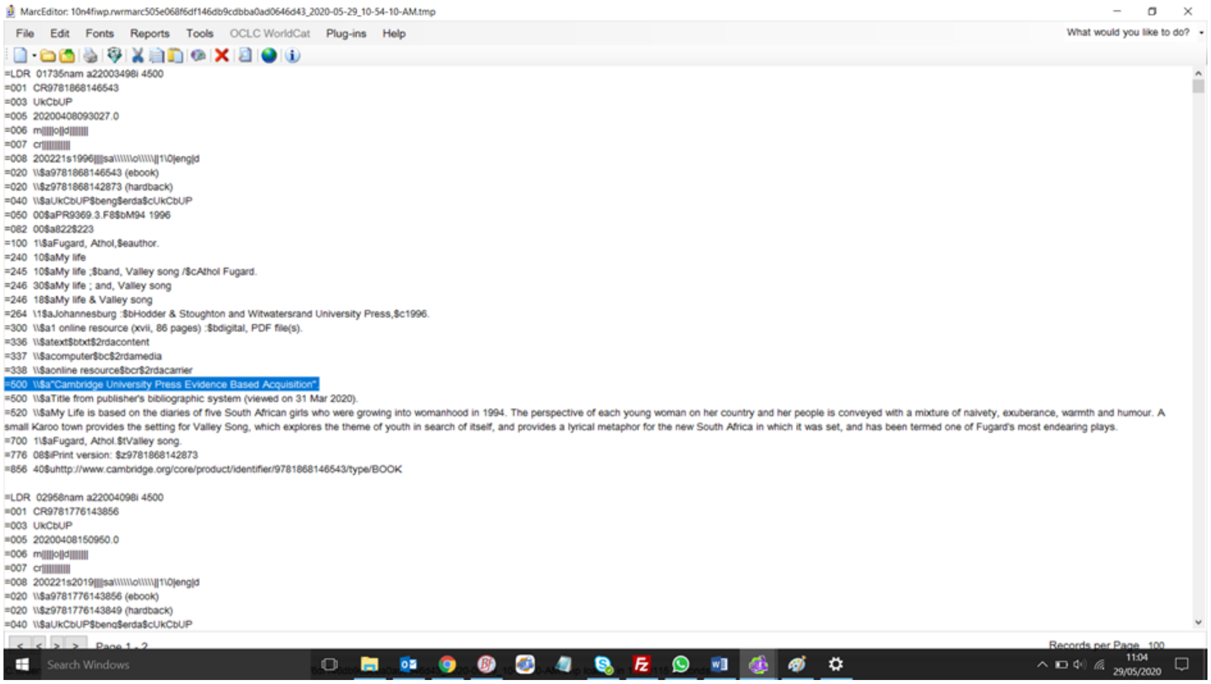
Data manipulation using MarcEdit

Concetta La Spada, Senior Metadata Librarian, Cambridge University Press
Twitter: @ConcettaLaSpada

One of my duties, as Senior Metadata Librarian at Cambridge University Press (CUP), is to modify records according to librarians' requests. It is one of the most interesting of my duties.

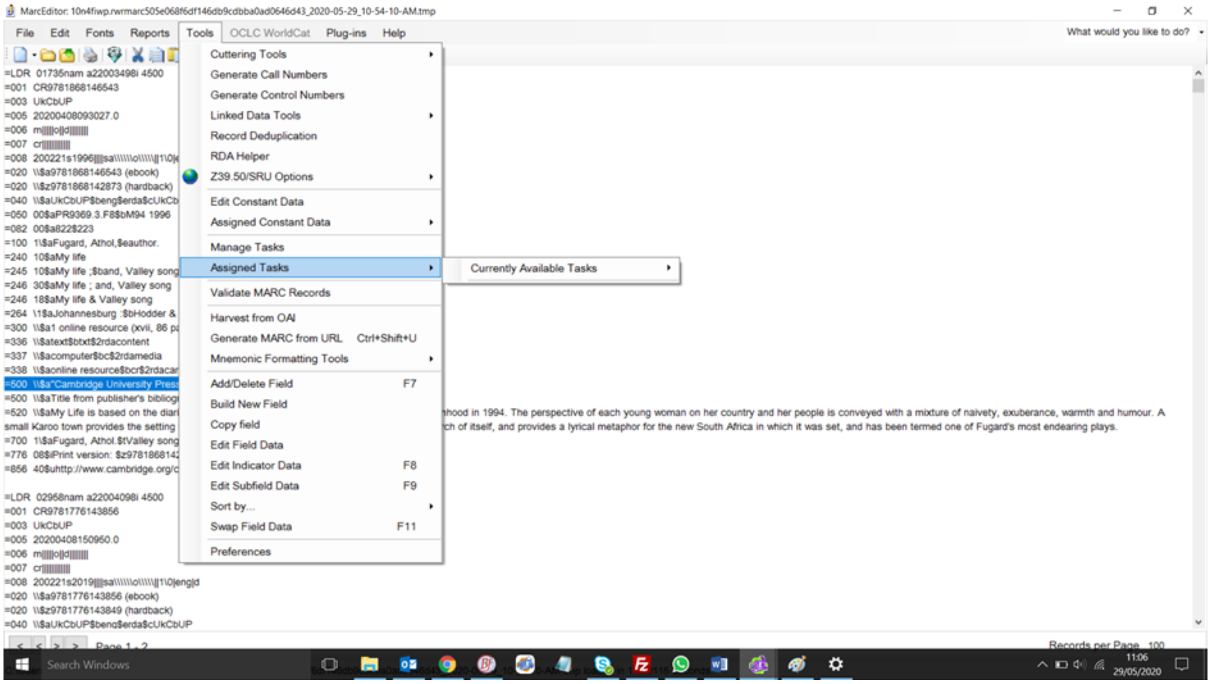
And with MarcEdit, the data manipulation program created by Terry Reese, it can be done quite easily.

The below screenshot is of a set of MARC records which I have manipulated. The library wanted all the records that they received from CUP to be easily identified. Their deal with us is called Evidence-Based Acquisition (EBA). Through this deal libraries can have access to our collections (full, or specific collections only) for a time, usually six months or one year, and after that decide whether to continue with the deal or purchase the titles that have been used the most by their users. At the end of this period, if they don't continue with the deal, MARC records will have to be removed from their online catalogues. By adding a special field to identify the batch belonging to the EBA deal with CUP, the removal process will be much easier.


MarcEditor: 10n4fiwp.rwrmarc505e068f6df146db9cdbba0ad0646d43_2020-05-29_10-54-10-AM.tmp
File
Edit
Fonts
Reports
Tools
OCLC WorldCat
Plug-ins
Help
What would you like to do?
Page 1 - 2
Records per Page 100
Search Windows
11:04
29/05/2020

```
=LDR  01735nam a22003498i 4500
=001  CR9781868146543
=003  UkCbUP
=005  20200408093027.0
=006  m|||||o||d||||||||
=007  cr|||||||||||
=008  200221s1996||||sa\\\\\\o\\\\\||1\0|eng|d
=020  \\$a9781868146543 (ebook)
=020  \\$z9781868142873 (hardback)
=040  \\$aUkCbUP$beng$erda$cUkCbUP
=050  00$aPR9369.3.F8$bM94 1996
=082  00$a822$223
=100  1\$aFugard, Athol,$eauthor.
=240  10$aMy life
=245  10$aMy life ;$band, Valley song /$cAthol Fugard.
=246  30$aMy life ; and, Valley song
=246  18$aMy life & Valley song
=264  \1$aJohannesburg :$bHodder & Stoughton and Witwatersrand University Press,$c1996.
=300  \\$a1 online resource (xvii, 86 pages) :$bdigital, PDF file(s).
=336  \\$atext$btxt$2rdacontent
=337  \\$acomputer$bc$2rdamedia
=338  \\$aonline resource$bcr$2rdacarrier
=500  \\$a"Cambridge University Press Evidence Based Acquisition".
=500  \\$aTitle from publisher's bibliographic system (viewed on 31 Mar 2020).
=520  \\$aMy Life is based on the diaries of five South African girls who were growing into womanhood in 1994. The perspective of each young woman on her country and her people is conveyed with a mixture of naivety, exuberance, warmth and humour. A
small Karoo town provides the setting for Valley Song, which explores the theme of youth in search of itself, and provides a lyrical metaphor for the new South Africa in which it was set, and has been termed one of Fugard's most endearing plays.
=700  1\$aFugard, Athol.$tValley song.
=776  08$iPrint version: $z9781868142873
=856  40$uhttp://www.cambridge.org/core/product/identifier/9781868146543/type/BOOK

=LDR  02958nam a22004098i 4500
=001  CR9781776143856
=003  UkCbUP
=005  20200408150950.0
=006  m|||||o||d||||||||
=007  cr|||||||||||
=008  200221s2019||||sa\\\\\\o\\\\\||1\0|eng|d
=020  \\$a9781776143856 (ebook)
=020  \\$z9781776143849 (hardback)
=040  \\$aUkCbUP$beng$erda$cUkCbUP
```

For every library who has asked for this kind of manipulation within their records, I have created "tasks" within MarcEdit, so that the process has become a sort of automated one. I simply click on the corresponding task and, in a few seconds, fields are removed and modified accordingly to what is requested. The first time I used the program, it gave me the feeling that I had real control of the data. I could use it in every way I wished. In fact, nowadays, I don't think about the MARC records I create (my main duty is to catalogue our eBooks, with new ones published every month) as a single kind of metadata. I feel more aware that the data I am putting together could potentially be used for other metadata formats. It makes me even more conscious that the good quality of it must be my utmost concern. Otherwise, any mistake will multiply the more that data is used.



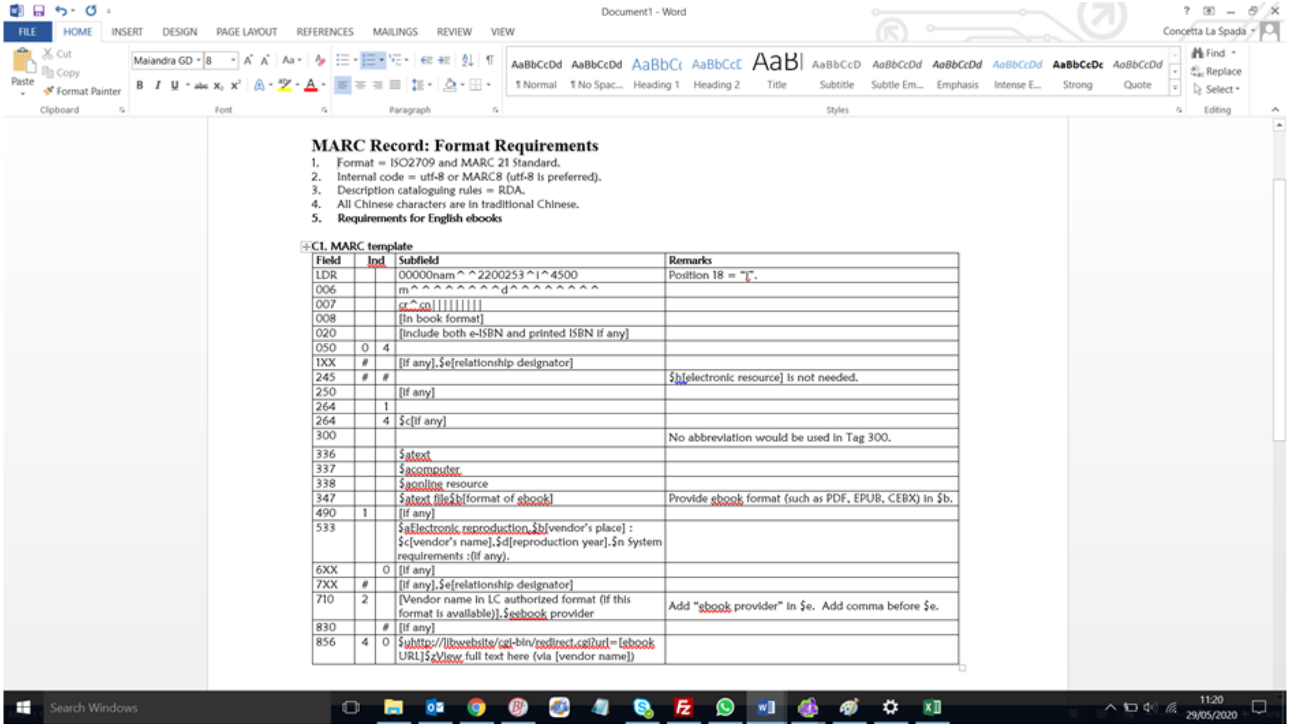
```
=LDR  01735nam a22003498i 4500
=001  CR9781868146543
=003  UkCbUP
=005  20200408093027.0
=006  m|||||o||d||||||||
=007  cr||||||||||||
=008  200221s1996||||sa\\\\\\o\\\\\||1\0|e
=020  \\$a9781868146543 (ebook)
=020  \\$z9781868142873 (hardback)
=040  \\$aUkCbUP$beng$erda$cUkCb
=050  00$aPR9369.3.F8$bM94 1996
=082  00$a822$223
=100  1\$aFugard, Athol,$eauthor.
=240  10$aMy life
=245  10$aMy life ;$band, Valley song
=246  30$aMy life ; and, Valley song
=246  18$aMy life & Valley song
=264  \1$aJohannesburg :$bHodder &
=300  \\$a1 online resource (xvii, 86 p
=336  \\$atext$btxt$2rdacontent
=337  \\$acomputer$bc$2rdamedia
=338  \\$aonline resource$bcr$2rdacar
=500  \\$a"Cambridge University Press
=500  \\$aTitle from publisher's bibliog
=520  \\$aMy Life is based on the diari
hood in 1994. The perspective of each young woman on her country and her people is conveyed with a mixture of naivety, exuberance, warmth and humour. A
small Karoo town provides the setting
ch of itself, and provides a lyrical metaphor for the new South Africa in which it was set, and has been termed one of Fugard's most endearing plays.
=700  1\$aFugard, Athol.$tValley song
=776  08$iPrint version: $z978186814
=856  40$uhttp://www.cambridge.org/c

=LDR  02958nam a22004098i 4500
=001  CR9781776143856
=003  UkCbUP
=005  20200408150950.0
=006  m|||||o||d||||||||
=007  cr||||||||||||
=008  200221s2019||||sa\\\\\\o\\\\\||1\0|eng|d
=020  \\$a9781776143856 (ebook)
=020  \\$z9781776143849 (hardback)
=040  \\$aUkCbUP$beng$erda$cUkCbUP
```

In a way, the manipulation I have illustrated above was quite easy.

Specific requests all together

A more intriguing scenario is when there are multiple fields to be removed or modified.

The below case is the case of a library who asked us for the following specification:



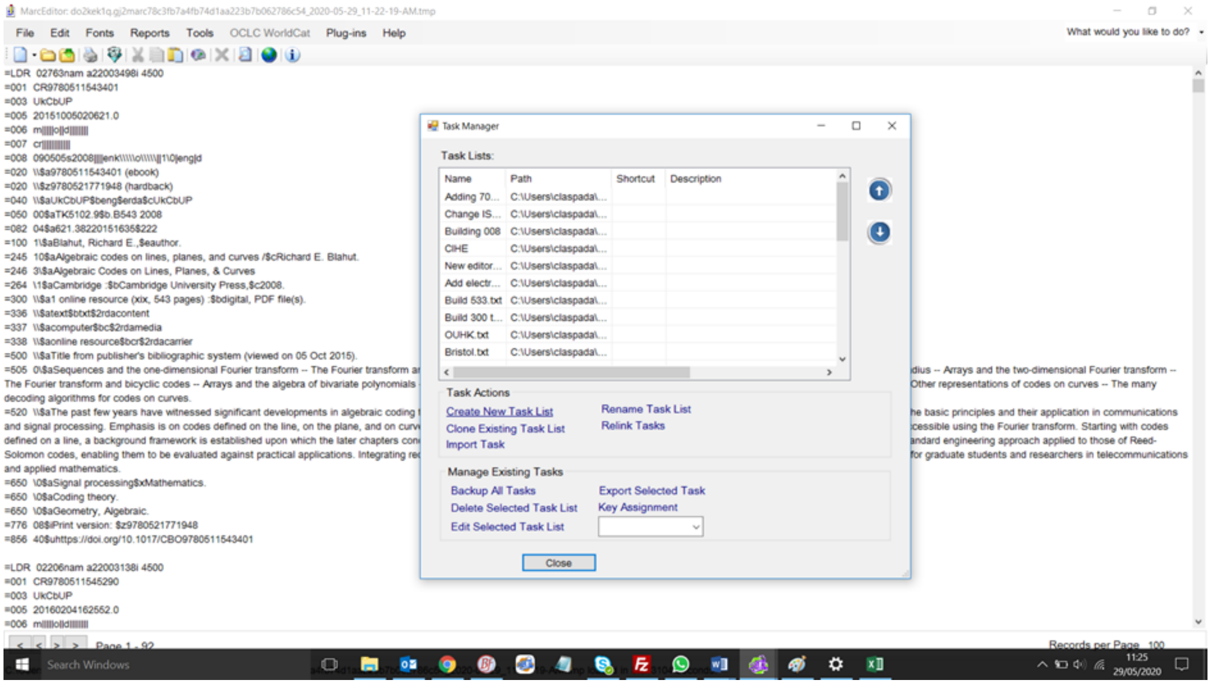
MARC Record: Format Requirements

1. Format = ISO2709 and MARC 21 Standard.
2. Internal code = utf-8 or MARC8 (utf-8 is preferred).
3. Description cataloguing rules = RDA.
4. All Chinese characters are in traditional Chinese.
5. **Requirements for English ebooks**

C1. MARC template

Field	Ind		Subfield	Remarks
LDR			00000nam^^2200253^i^4500	Position 18 = "i".
006			m^^^^^^^^^d^^^^^^^^^	
007			cr^cn|||||||||	
008			[In book format]	
020			[include both e-ISBN and printed ISBN if any]	
050	0	4		
1XX	#		[If any].$e[relationship designator]	
245	#	#		$h[electronic resource] is not needed.
250			[If any]	
264		1		
264		4	$c[If any]	
300				No abbreviation would be used in Tag 300.
336			$atext	
337			$acomputer	
338			$aonline resource	
347			$atext file$b[format of ebook]	Provide ebook format (such as PDF, EPUB, CEBX) in $b.
490	1		[If any]	
533			$aElectronic reproduction.$b[vendor's place] : $c[vendor's name],$d[reproduction year].$n System requirements :(If any).	
6XX		0	[If any]	
7XX	#		[If any].$e[relationship designator]	
710	2		[Vendor name in LC authorized format (If this format is available)].$eebook provider	Add "ebook provider" in $e. Add comma before $e.
830		#	[If any]	
856	4	0	$uhttp://libwebsite/cgi-bin/redirect.cgi?url=[ebook URL]$zView full text here (via [vendor name])	

When I saw it first I thought that it would have required careful planning, but that it would be very interesting to be able to provide something like this. In MarcEdit, after opening the set of records that has to be modified, we click on Tools – Manage Tasks – Create new tasks:



MarcEditor: do2kek1q.gj2marc78c3fb7a4fb74d1aa223b7b062786c54_2020-05-29_11-22-19-AM.tmp

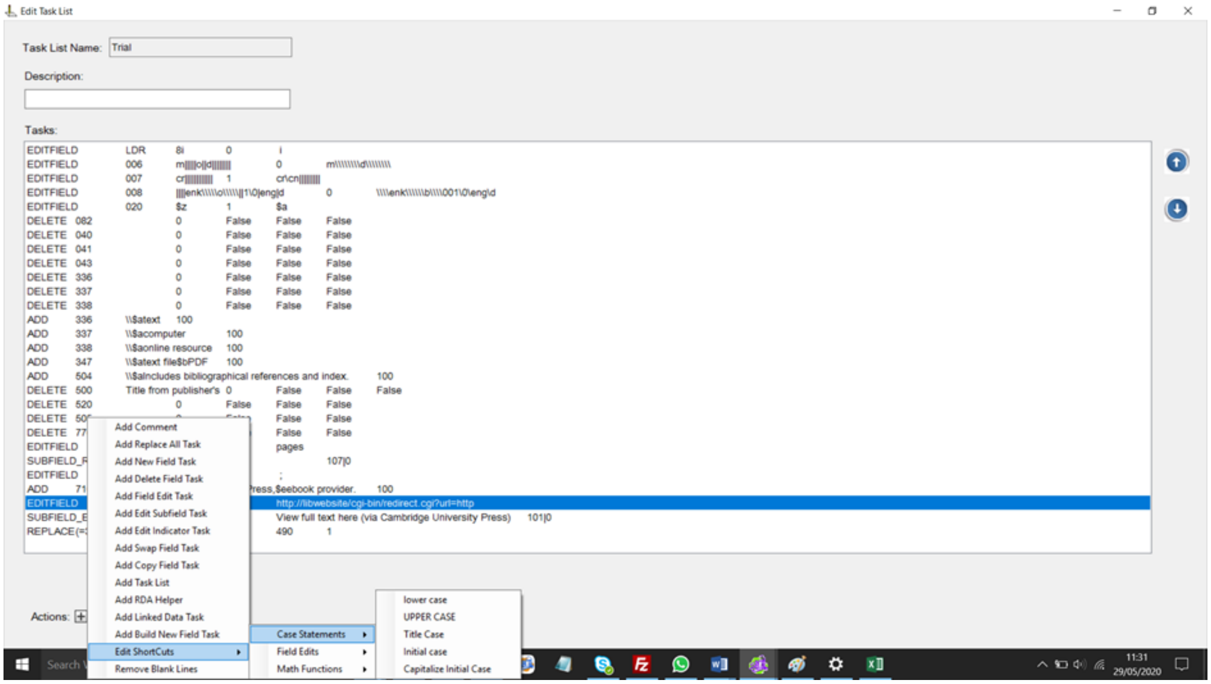
```
=LDR  02763nam a22003498i 4500
=001  CR9780511543401
=003  UkCbUP
=005  20151005020621.0
=006  m|||||o||d||||||||
=007  cr||||||||||||
=008  090505s2008||||enk\\\\\o\\\\\||1\0|eng|d
=020  \\$a9780511543401 (ebook)
=020  \\$z9780521771948 (hardback)
=040  \\$aUkCbUP$beng$erda$cUkCbUP
=050  00$aTK5102.9$b.B543 2008
=082  04$a621.382201516 35$222
=100  1\$aBlahut, Richard E.,$eauthor.
=245  10$aAlgebraic codes on lines, planes, and curves /$cRichard E. Blahut.
=246  3\$aAlgebraic Codes on Lines, Planes, & Curves
=264  \1$aCambridge :$bCambridge University Press,$c2008.
=300  \\$a1 online resource (xix, 543 pages) :$bdigital, PDF file(s).
=336  \\$atext$btxt$2rdacontent
=337  \\$acomputer$bc$2rdamedia
=338  \\$aonline resource$bcr$2rdacarrier
=500  \\$aTitle from publisher's bibliographic system (viewed on 05 Oct 2015).
=505  0\$aSequences and the one-dimensional Fourier transform -- The Fourier transform a      dius -- Arrays and the two-dimensional Fourier transform --
The Fourier transform and bicyclic codes -- Arrays and the algebra of bivariate polynomials      Other representations of codes on curves -- The many
decoding algorithms for codes on curves.
=520  \\$aThe past few years have witnessed significant developments in algebraic coding      he basic principles and their application in communications
and signal processing. Emphasis is on codes defined on the line, on the plane, and on curv      cessible using the Fourier transform. Starting with codes
defined on a line, a background framework is established upon which the later chapters con      andard engineering approach applied to those of Reed-
Solomon codes, enabling them to be evaluated against practical applications. Integrating re      for graduate students and researchers in telecommunications
and applied mathematics.
=650  \0$aSignal processing$xMathematics.
=650  \0$aCoding theory.
=650  \0$aGeometry, Algebraic.
=776  08$iPrint version: $z9780521771948
=856  40$uhttps://doi.org/10.1017/CBO9780511543401

=LDR  02206nam a22003138i 4500
=001  CR9780511545290
=003  UkCbUP
=005  20160204162552.0
=006  m|||||o||d||||||||
```

Task Manager

Task Lists:

Name	Path	Shortcut	Description
Adding 70...	C:\Users\claspada\...		
Change IS...	C:\Users\claspada\...		
Building 008	C:\Users\claspada\...		
CIHE	C:\Users\claspada\...		
New editor...	C:\Users\claspada\...		
Add electr...	C:\Users\claspada\...		
Build 533.txt	C:\Users\claspada\...		
Build 300 t...	C:\Users\claspada\...		
OUHK.txt	C:\Users\claspada\...		
Bristol.txt	C:\Users\claspada\...		

Even in a case with as many modifications as this one, there weren't any particularly difficulties because the program provides many options to choose from. In this way, it is possible to create any kind of record, according to specific requests.



Edit Task List

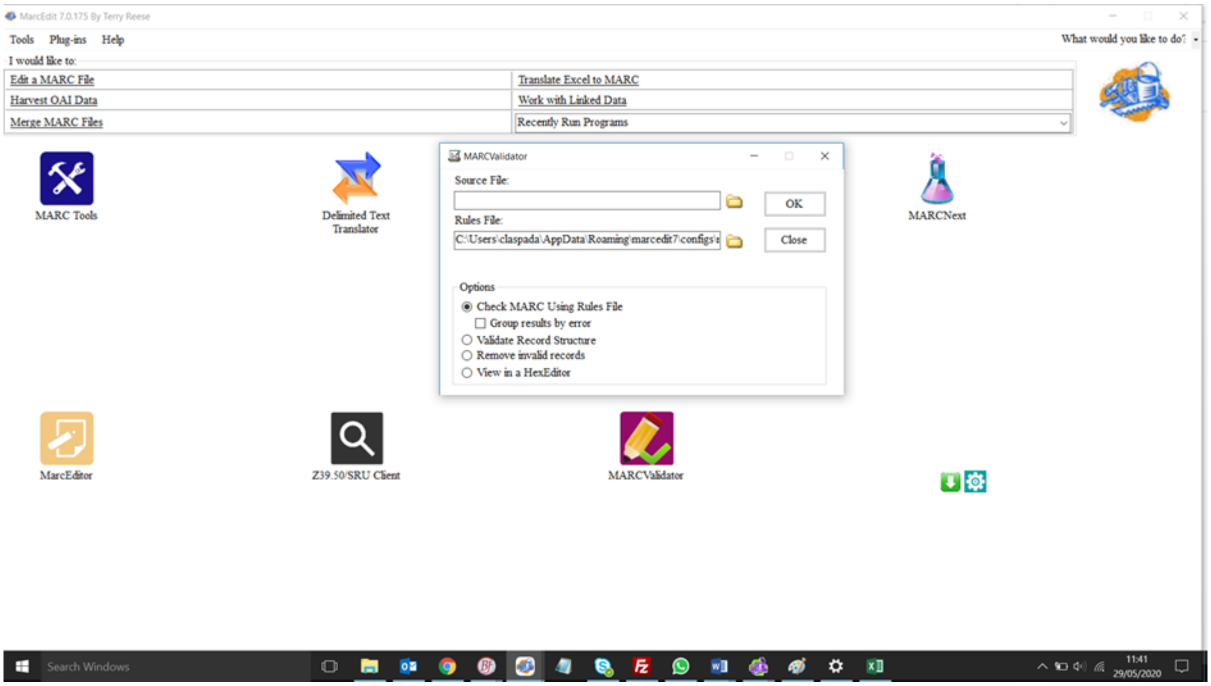
Task List Name: Trial

Description:

Tasks:

```
EDITFIELD     LDR    8i            0              i
EDITFIELD     006    m|||||o||d||||||||           0          m\\\\\\\\d\\\\\\\\
EDITFIELD     007    cr||||||||||||    1          cr\cn||||||||||
EDITFIELD     008    ||||enk\\\\\o\\\\\||1\0|eng|d          0         \\\\enk\\\\\\b\\\\001\0\eng\d
EDITFIELD     020    $z            1              $a
DELETE   082         0             False     False     False
DELETE   040         0             False     False     False
DELETE   041         0             False     False     False
DELETE   043         0             False     False     False
DELETE   336         0             False     False     False
DELETE   337         0             False     False     False
DELETE   338         0             False     False     False
ADD      336    \\$atext    100
ADD      337    \\$acomputer        100
ADD      338    \\$aonline resource     100
ADD      347    \\$atext file$bPDF    100
ADD      504    \\$aIncludes bibliographical references and index.       100
DELETE   500    Title from publisher's 0     False     False     False
DELETE   520         0             False     False     False
DELETE   50                                  False     False
DELETE   77                                  False     False
EDITFIELD                                    pages
SUBFIELD_R                                             107|0
EDITFIELD                                     ;
ADD      71                            ress,$eebook provider.       100
EDITFIELD                                    http://libwebsite/cgi-bin/redirect.cgi?url=http
SUBFIELD_E                                   View full text here (via Cambridge University Press)     101|0
REPLACE(=                                    490       1
```

And it doesn’t matter how many records there are in a single batch. The process will require only seconds for small batches, and a few minutes for big ones. It is very important, though, to make sure that the changes made do not enter into conflict with anything else. I am thinking specifically of the 008 field: “Field 008 contains 40 character positions (00-39) that provide coded information about the record as a whole and about special bibliographic aspects of the item being catalogued. These coded data elements are potentially useful for retrieval and data management purposes.”[1] So it is fundamental that, when this field is modified, like in the case I have illustrated, any modification does not alter its structure in an incorrect way. For this reason I always check MARC records batches with a very useful tool. The MARC Validator:



Source File:

Rules File: C:\Users\claspada\AppData\Roaming\marcedit7\configs\

Options
● Check MARC Using Rules File
□ Group results by error
○ Validate Record Structure
○ Remove invalid records
○ View in a HexEditor

This tool picks up any errors, cataloguing or structural, within the selected batch and gives them back in a list, so that cataloguers can see which records are affected – and how – and can modify the “Task” they have created accordingly. Being the only one doing metadata manipulation of this kind, I do not have anyone else who can check my work, so this tool is fundamental to me.

The importance of knowing cataloguing rules and MARC 21 format

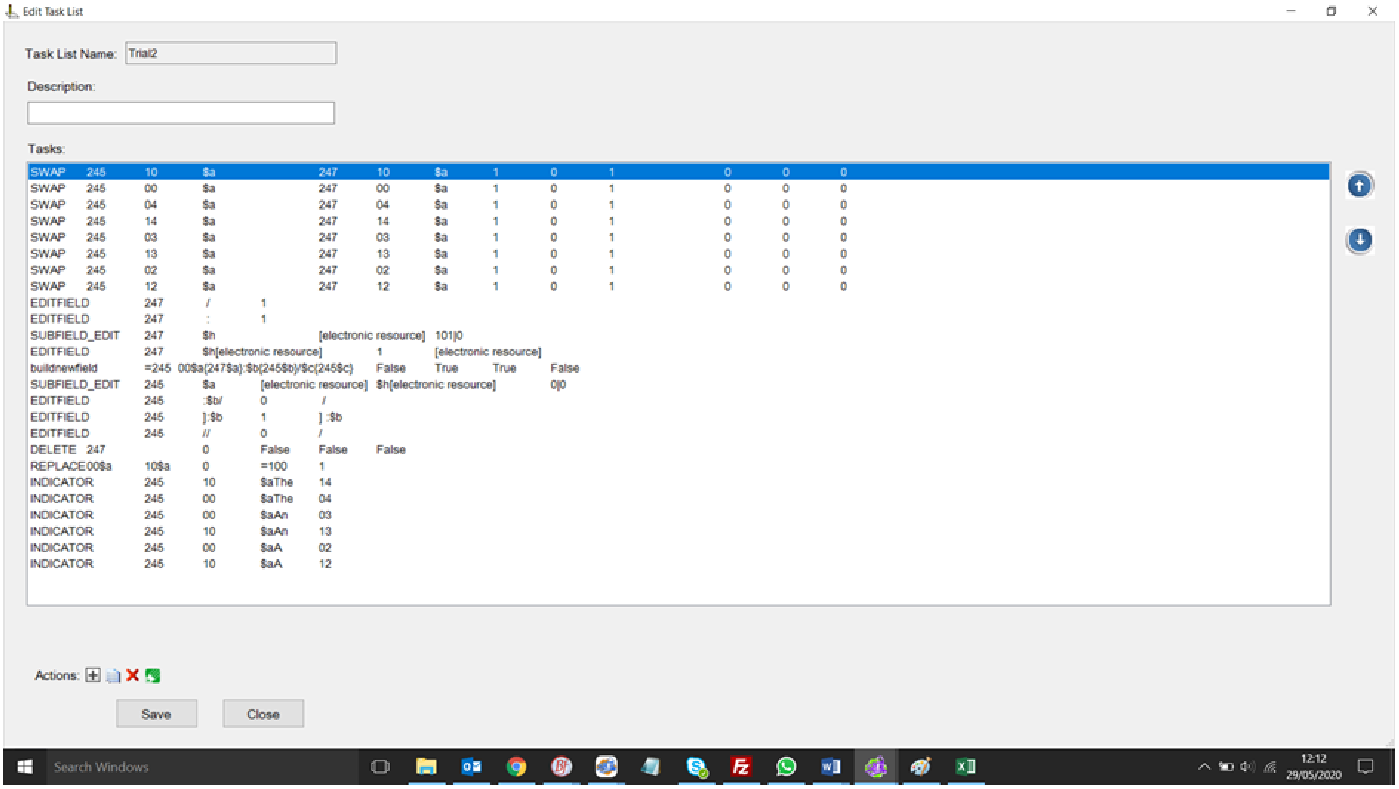
A case that proves how, in order to modify records according to cataloguing rules, the person creating the task has to have a deep knowledge of cataloguing rules, and of the MARC 21 format’s structure, is the following one. A library asked that all their records had the subfield $h (“Medium” subfield) in every 245 field (Title field) with the data [electronic resource].

You may think, ok, it is just one subfield, what is the difficulty?

[1] *Cataloger’s Reference Shelf* by The Library Corporation, https://www.itsmarc.com/crs/crs.htm

Well, the first thing is that the subfield $h has to be “attached” to the last word of the subfield $a (“Main title” subfield) with no spaces. In order to solve this I needed to take all the data from the 245 field, in every record of course, and use it to create a new 245 field that would include a subfield $h. First I had to copy the data of the 245 subfield $a into a new field, 247, including the indicators; then to add the subfield $h to it, and then to create a new 245 field. This new field would be composed of the 247’s subfield $a, and the subfields $c (“Statement of responsibility” subfield) and $b (“Subtitle” subfield) of the original 245 field.

Following this step the 247 field could be deleted as it was no longer necessary. An important step to remember is to always check at the end of such a manipulation (and of any manipulation, really) that any of the fields that have been added temporarily have been removed. We do not want them to appear in the final “product”.



Task List Name: Trial2

Description:

Tasks:

```
SWAP    245   10     $a                       247    10     $a     1      0      1             0      0      0
SWAP    245   00     $a                       247    00     $a     1      0      1             0      0      0
SWAP    245   04     $a                       247    04     $a     1      0      1             0      0      0
SWAP    245   14     $a                       247    14     $a     1      0      1             0      0      0
SWAP    245   03     $a                       247    03     $a     1      0      1             0      0      0
SWAP    245   13     $a                       247    13     $a     1      0      1             0      0      0
SWAP    245   02     $a                       247    02     $a     1      0      1             0      0      0
SWAP    245   12     $a                       247    12     $a     1      0      1             0      0      0
EDITFIELD     247     /        1
EDITFIELD     247     :        1
SUBFIELD_EDIT 247    $h                       [electronic resource]  101|0
EDITFIELD     247    $h[electronic resource]        1      [electronic resource]
buildnewfield =245  00$a[247$a]:$b[245$b]/$c[245$c]  False  True   True   False
SUBFIELD_EDIT 245    $a       [electronic resource]  $h[electronic resource]     0|0
EDITFIELD     245    :$b/     0        /
EDITFIELD     245    ]:$b     1        ] :$b
EDITFIELD     245    //       0        /
DELETE  247          0        False    False  False
REPLACE00$a   10$a   0        =100     1
INDICATOR     245    10       $aThe    14
INDICATOR     245    00       $aThe    04
INDICATOR     245    00       $aAn     03
INDICATOR     245    10       $aAn     13
INDICATOR     245    00       $aA      02
INDICATOR     245    10       $aA      12
```

The work could not end here though. I had to take into account that there are titles that do not have subtitles (so no subfield b), so I had to add an “action” that would delete the text “:$b”, which would appear in any new 245 field even if there was no data following it.

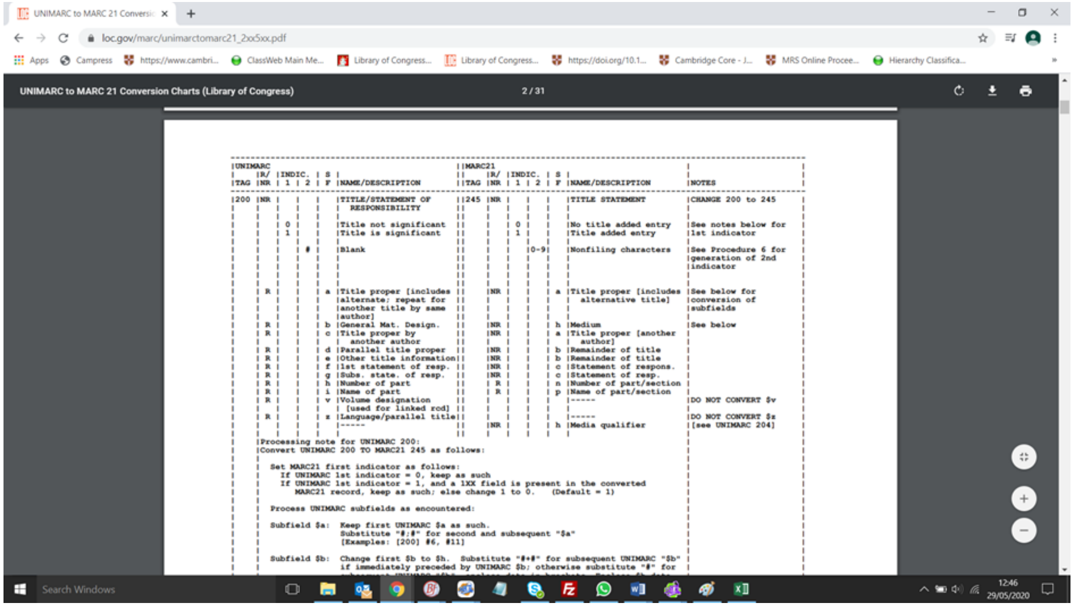
Also, since by adding a new 245 I could not give specific indicators, and I could not use the ones I had copied, I had to make sure that every 245 field beginning with “The”, and with indicators 1 and 0, had the correct indicators 1 and 4, and so on for all the other cases: 0 and 4 for those 245 fields beginning with “The” where there was not a 1xx field; 0 and 3 for those 245 fields beginning with “An” where there was not a 1xx field; and so on (see screenshot above).

You can see, from the above example, why, to use this program effectively, the person using it has to know cataloguing rules, and be aware of all the scenarios that can be encountered. If I had not known the rules related to use of indicators in cases where titles begins with articles, I would have created incorrect metadata. It is important to always think ahead. To think also about human error, which lurks around every corner. This is something that only experience can teach.

I know nowadays, for example, not to trust the data I deal with 100%, because that data has been created or modified manually at some point in its “journey”, and human error can happen any time.

From one format to another

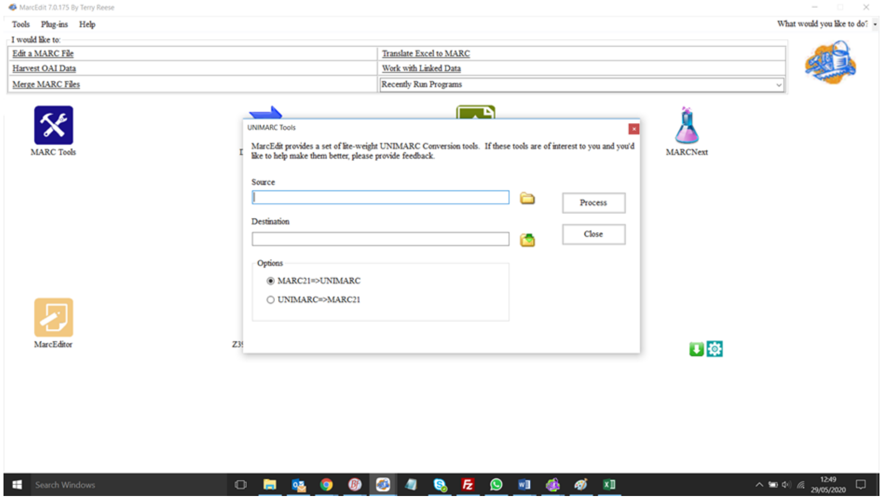
As we know, MARC 21 is the format most used among libraries worldwide. However, there are institutions which use other MARC formats. The one I come across the most is UNIMARC.



```
-------------------------------------------------------------------------------------------------
|UNIMARC                                       ||MARC21                                  |                    |
|    |R/ |INDIC. | S |                         ||    |R/ |INDIC. | S |                   |                    |
|TAG |NR | 1 | 2 | F |NAME/DESCRIPTION         ||TAG |NR | 1 | 2 | F |NAME/DESCRIPTION   |NOTES               |
-------------------------------------------------------------------------------------------------
|200 |NR |   |   |   |TITLE/STATEMENT OF       ||245 |NR |   |   |   |TITLE STATEMENT    |CHANGE 200 to 245   |
|    |   |   |   |   |  RESPONSIBILITY         ||    |   |   |   |   |                   |                    |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |                    |
|    |   | 0 |   |   |Title not significant    ||    |   | 0 |   |   |No title added entry|See notes below for |
|    |   | 1 |   |   |Title is significant     ||    |   | 1 |   |   |Title added entry  |1st indicator       |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |                    |
|    |   |   | # |   |Blank                    ||    |   |   |0-9|   |Nonfiling characters|See Procedure 6 for |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |generation of 2nd   |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |indicator           |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |                    |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |                    |
|    | R |   |   | a |Title proper [includes   ||    |NR |   |   | a |Title proper [includes|See below for    |
|    |   |   |   |   |alternate; repeat for    ||    |   |   |   |   |  alternative title]|conversion of      |
|    |   |   |   |   |another title by same    ||    |   |   |   |   |                   |subfields           |
|    |   |   |   |   |author]                  ||    |   |   |   |   |                   |                    |
|    | R |   |   | b |General Mat. Design.     ||    |NR |   |   | h |Medium             |See below           |
|    | R |   |   | c |Title proper by          ||    |NR |   |   | a |Title proper [another|                  |
|    |   |   |   |   |  another author         ||    |   |   |   |   |  author]          |                    |
|    | R |   |   | d |Parallel title proper    ||    |NR |   |   | b |Remainder of title |                    |
|    | R |   |   | e |Other title information  ||    |NR |   |   | b |Remainder of title |                    |
|    | R |   |   | f |1st statement of resp.   ||    |NR |   |   | c |Statement of respons.|                  |
|    | R |   |   | g |Subs. state. of resp.    ||    |NR |   |   | c |Statement of resp. |                    |
|    | R |   |   | h |Number of part           ||    | R |   |   | n |Number of part/section|                 |
|    | R |   |   | i |Name of part             ||    | R |   |   | p |Name of part/section|                   |
|    | R |   |   | v |Volume designation       ||    |   |   |   |   |-----              |DO NOT CONVERT $v   |
|    |   |   |   |   |  [used for linked rcd]  ||    |   |   |   |   |                   |                    |
|    | R |   |   | z |Language/parallel title  ||    |   |   |   |   |-----              |DO NOT CONVERT $z   |
|    |   |   |   |   |-----                    ||    |NR |   |   | h |Media qualifier    |[see UNIMARC 204]   |
|    |   |   |   |   |                         ||    |   |   |   |   |                   |                    |
|    |Processing note for UNIMARC 200:                                                   |                    |
|    |Convert UNIMARC 200 TO MARC21 245 as follows:                                      |                    |
|    |                                                                                   |                    |
|    |  Set MARC21 first indicator as follows:                                           |                    |
|    |    If UNIMARC 1st indicator = 0, keep as such                                     |                    |
|    |    If UNIMARC 1st indicator = 1, and a 1XX field is present in the converted      |                    |
|    |       MARC21 record, keep as such; else change 1 to 0.    (Default = 1)           |                    |
|    |                                                                                   |                    |
|    |  Process UNIMARC subfields as encountered:                                        |                    |
|    |                                                                                   |                    |
|    |  Subfield $a:  Keep first UNIMARC $a as such.                                     |                    |
|    |                Substitute "#;#" for second and subsequent "$a"                    |                    |
|    |                [Examples: [200] #6, #11]                                          |                    |
|    |                                                                                   |                    |
|    |  Subfield $b:  Change first $b to $h.  Substitute "#+#" for subsequent UNIMARC "$b"|                   |
|    |                if immediately preceded by UNIMARC $b; otherwise substitute "#" for |                  |
```

A European library asked us if we could send them UNIMARC records. I had used this format at the beginning of my career, when I first started cataloguing books, and I knew already that there are important differences between these two formats. An example is the title field, which in MARC 21 is the 245 field, and in UNIMARC is the 200 field.

Fortunately, the Library of Congress provides mapping documents, which show the differences between the two formats. It would have been an very long task to do by myself though, with too many unknowns; also, as I had not worked with UNIMARC for years I could have missed something. Fortunately, MarcEdit has a tool which converts MARC21 records to UNIMARC records:



MarcEdit provides a set of lite-weight UNIMARC Conversion tools. If these tools are of interest to you and you'd like to help make them better, please provide feedback.

● MARC21=>UNIMARC
○ UNIMARC=>MARC21

All I had to do was to click on “Process” and the records were changed to the UNIMARC format. But checking the file, I noticed that the 856 field, the link field, had been deleted. This was because there was no corresponding UNIMARC field. The documents from the Library of Congress did not give a corresponding field, probably because at that time electronic resources were not much used. At the time I just added it back, by merging the new records with the old ones, and the library which had requested this format confirmed that the records were good and could be used. However, I wanted to make this issue known, so I wrote to Terry Reese and he found new guidelines from IFLA that specified that the 856 field was the URL field even within the UNIMARC format. So the tool was updated accordingly. MarcEdit, I found, is very much a work in progress, and I find this one of its best features. It can be updated and modified, accordingly to what users want to do with it.

Conclusion

The work that Terry Reese has done is priceless. To create such a program, update it, and make it available free of charge to anyone is amazing. There are many programs and databases, within the library sector, that are not available free of charge, and even if I understand the reasons for it, I still feel that it would be so much better if they were. There are many libraries that do not have the money for such expenses, even though their librarians could achieve so much more if they had access to these resources.

The above cases are only a few examples of what can be done with MarcEdit. I hope that what I have written will encourage many to use this program and to “have fun” with it. The possibilities are endless!

How Library Carpentry helps us to transform our data and our training

Phil Reed, Data Specialist (Teaching, Learning and Students), The University of Manchester Library

There is much discussion about the importance of digital skills requirements across the sector, involving groups such as RLUK Digital Scholarship Network, The National Archives, and Jisc Digital Capabilities. Such matters are particularly relevant for metadata and discovery (Daniels, 2020). The issues include establishing learners' requirements, developing lessons and competency frameworks, and sourcing an affordable and sustainable pool of instructors. This paper will look at the delivery of Library Carpentry workshops in Manchester and Durham, how Library Carpentry can help to provide an opportunity to develop a community of good practice, and how these are relevant to metadata and cataloguing work. It ends with two case studies: what are people inspired to do after attending Library Carpentry workshops, and the development of a Carpentries-style lesson at Manchester about TEI XML (Text Encoding Initiative Extensible Markup Language).

Introducing Library Carpentry to Manchester

Staff at The University of Manchester Library (UML) have been inspired by the lessons and workshops presented by Library Carpentry (LC). In short, Library Carpentry teaches information professionals how to apply best practice when transforming and working with data sets; it is a non-profit organisation and an international community of volunteers who help people without a background in programming. In the last few years, three UML staff have qualified as LC instructors (Carlene Barton, Nilani Ganeshwaran and Phil Reed). They ran a workshop in November 2018 for their colleagues and those at nearby Manchester Metropolitan University, teaching good practice with tidy data and an introduction to OpenRefine. Feedback was positive and constructive; after one workshop, participants came away with skills and were eager to apply them (although some reported that they would not be able to use the skills right away due to commitments elsewhere).

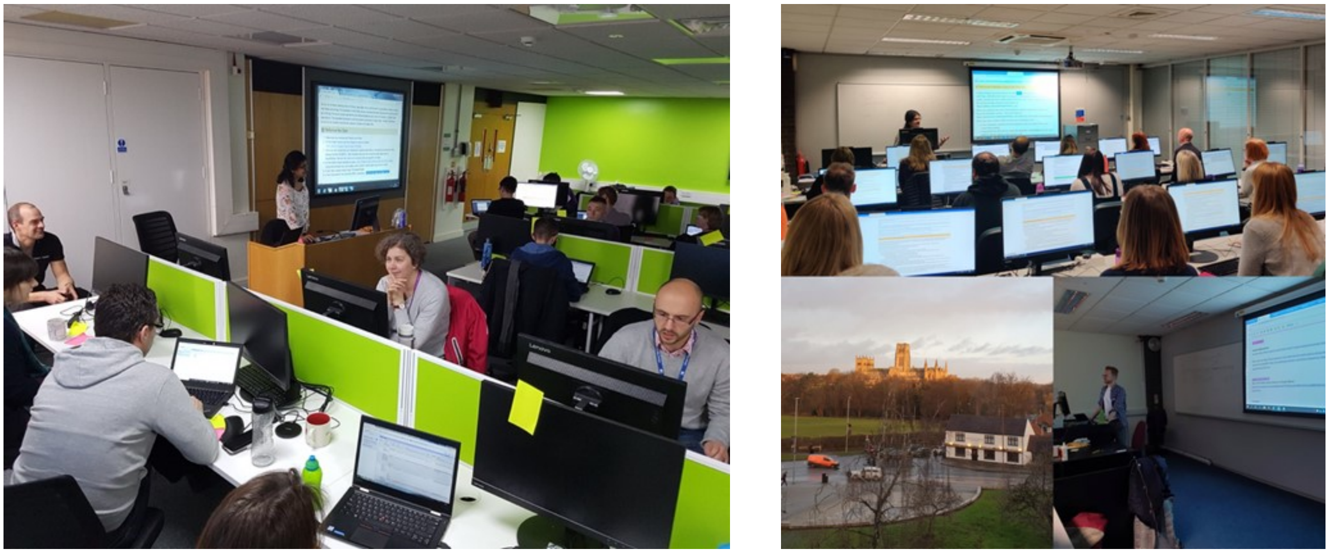


Figure 1: Library Carpentry at Manchester, 2018 (left) and Durham, 2019 (right)

Taking the show on the road

In December 2019, the Manchester Library Carpentry team were invited to deliver a workshop for North East England. Kindly hosted by Durham University, the syllabus was slightly amended to cover the concept of regular expressions as well as OpenRefine. The icebreaker section was about jargon busting, asking people to share an acronym, name, or concept for the hosts or other participants to explain, such as Unicode or CRIS. Everyone in the room learned something new within the first 15 minutes and could be reassured that nobody knows everything. The workshop was wrapped up by asking learners to set goals for using their new skills; what will they do after two weeks, and after three months (see the case study below for more about these goals).

The feedback was more positive from the Durham session; more people found it immediately useful. This is likely due to the application process and demographic; a greater proportion of people who manipulate data sets very frequently in their job came to the Durham workshop.

"Eye opening! Amazing. This will be really useful for anyone using spreadsheets."

– feedback from a participant in Durham

Relevance to metadata and cataloguing librarians

There are many immediately relevant opportunities for librarians working with metadata and cataloguing, people for whom cleaning and transforming data are daily tasks. Learning OpenRefine is a way for people to do something immediately; for example, after attending the workshop, someone could open a CSV file and fix inconsistently formatted author names and initials. The Library Carpentry workshop lays a foundation for learning more computational processes; for example, one could learn how to write simple Python scripts to fix inconsistent formatting in a batch of files (other lessons which can be used in a longer Library Carpentry workshop are listed in a later section of this article).

Community is key

Library Carpentry is one of three communities within a parent group called The Carpentries. The other communities of Software Carpentry and Data Carpentry are more mature (dating back to 1998) and are usually taught to academic researchers and students. All lesson content across The Carpentries has been developed by volunteers and is available under a Creative Commons Attribution licence. Workshops are delivered by volunteers and can be arranged centrally (for a fee) or by one of the 80 member organisations from 10 countries around the world. Workshops largely follow a set pattern or syllabus and must be taught by at least one qualified Carpentries instructor. All necessary guidance is provided in the handbook, including the workshop Code of Conduct which ensures a safe and comfortable learning environment (The Carpentries, 2020).

There are sound, evidence-based pedagogical practices applied to all Carpentries lessons, taught to all instructors before they qualify (Allen et al., 2019). For example, formative assessment is used frequently, dismissive language is not permitted, and mistakes are used as learning opportunities. This quality foundation has helped in the delivery of 2,300 workshops run in 61 countries between 2012 and 2019.

Working with data is important in fields such as biology, where researchers and support staff are increasingly required to handle "big data" and apply "FAIR Principles" for data management[1] (Wilkinson et al., 2016). It has been shown that sharing good practices using the Carpentries model will "foster open, transparent, and sound scientific results beneficial to society" and that developing a community of practice "will bring many benefits to its members and to their institution" (Stevens et al., 2018). The benefits of applying FAIR Data Principles when working with metadata have been clearly established (Ball, 2020). It follows, therefore, that developing a community of practice for metadata specialists using Library Carpentry is an opportunity well worth exploring.

Looking into the future

The Manchester Library Carpentry team would like to run more Library Carpentry workshops and inspire more people to join as instructors. In the short term, any workshops might have to run online instead of in person, so more time will be required to investigate how that could work most effectively.

The team would also like to encourage people to contribute to lesson development. Anybody can suggest edits to the lessons hosted on GitHub; see the 'Repository' links on the lessons page. Each repository has a list of 'Issues' which are discussions about any changes between maintainers and with the public.

In addition to the tidy data, regular expressions, and OpenRefine lessons, Library Carpentry has a core curriculum designed to run over two days (though you can split these into four parts or half-days). These lessons include the UNIX shell and an introduction to Git. The curriculum can be extended to include SQL, webscraping, Python and data for archivists, although some of this content is in the early stages of development. There is a standalone lesson titled 'Top 10 FAIR Data & Software Things' which may be of interest to those working in research libraries and many others. Further lessons at the earliest stages of development include MarcEdit, Wikidata, R and XML.

Finally, you may wish to use the principles of Carpentries lessons, or indeed the lesson templates, following the Creative Commons Attribution licence. There are examples of lesson templates being adapted and reused from Neurohackweek, 23 (Research Data) Things and The University of Manchester Library's Introduction to TEI (see case study 2 below).

Case study 1: What can people do after Library Carpentry?

At the end of the Durham workshop, learners were asked "Based on what you have learned today, what will you do differently in the next two weeks? Three months?". Some of the responses received were:

- "Process chain of Excel commands but in OpenRefine."
- "Catalogue data, although getting it in and out of the catalogue is a factor."
- "Harmonising the way data is presented across data sets, e.g. author names."
- "Find faulty terms in thesaurus that broke. Data recovery after corruption."
- "Learn more about APIs."
- "SCONUL stats, renewals."
- "RegEx to collate reports, who has been using services."
- "Survey data, cleaning up."

Learning how to automate or articulate one's work using scripts may save time in the long run, and also increase the transparency and accountability of workflows.

[1] "FAIR Principles" mean that digital assets are findable, accessible, interoperable, and reusable.

Case study 2: Introduction to TEI

'Introduction to TEI XML' was a series of face-to-face workshops at The University of Manchester Library delivered in November and December 2019. It was designed by Jane Gallagher, Elizabeth Gow, Dr Jo Edge, and Phil Reed, with special thanks to Professor David Denison, Nuria Yáñez-Bouza, Dr Giles Bergel, and Dr Christopher Ohge. It was aimed at Special Collections staff, cataloguers, and other library and information professionals. There was formative assessment throughout, aligning with the learning outcomes, and opportunities for collaborative learning.

After the workshops completed, the materials were stored on the local intranet, and were thus unavailable to those outside the initial group. To adapt the workshop into an open educational resource, the author followed the open practice exemplified by The Carpentries and turned the materials into a lesson hosted on GitHub. The resource was released with a Creative Commons Attribution Non-commercial licence. Some adaptation was required; however, the initial approach was a close match pedagogically to the Carpentries format so this was quite straightforward.

It is currently a functional 'alpha' release with various small adjustments required. One day, it could inspire the development of the suggested Library Carpentry XML workshop. If the release is shown to effective, the approach of building Carpentries-style lessons may be suggested for other forms of staff development at Manchester. Meanwhile, Software Carpentries lessons continue to be taught at Manchester about Python and the Unix shell. These workshops have been running for three years, showing the ongoing value of the Carpentries approach to researchers.

References

Allen, K., Dennis, T., and Otsuji, R. (2019) '*Creating effective workshops using The Carpentries inclusive pedagogy*,' University of California Digital Library Forum (UC DLFx).

Ball, A. (2020) 'Metadata for better data.' *Catalogue & Index*, (198) pp. 17–20.

Daniels, J. (2020) 'What's in a name? from CIG to MDG to CPD.' *Catalogue and Index*, (198) pp. 4–7.

Stevens, S. L. R., Kuzak, M., Martinez, C., Moser, A., Bleeker, P., and Galland, M. (2018) 'Building a local community of practice in scientific programming for life scientists.' *PLOS Biology*. NLM (Medline), 16(11) p. e2005561. https://doi.org/10.1371/journal.pbio.2005561

The Carpentries (2020) *The Carpentries Handbook*. [Online] [Accessed on 25th May 2020] https://docs.carpentries.org/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016) 'The FAIR Guiding Principles for scientific data management and stewardship.' *Scientific Data*. Nature Publishing Groups, 3(1) pp. 1–9.

Book review: *Sudden position guide to cataloguing and metadata*

Reviewed by Liz Antell, Resource Assistant, Edinburgh Napier University

Sudden position guide to cataloguing and metadata **edited by Jeremy Myntti; contributors: Ben Abrahamse, Whitney Buccicone, Stephen Buss, Autumn Falkner, Matthew Gallagher, Jeremy Myntti, Nicole Smeltekop.**
Chicago: Association for Library Collections & Technical Services, a division of the American Library Association, 2019. 167 pages.
ISBN: 9780838948576

"So You're Suddenly a Cataloger", the first chapter greets you, in a manner reminiscent of the *Hitchhiker's Guide to the Galaxy* advising you, in large friendly letters, DON'T PANIC.

The book is reassuring and welcoming from the outset. Targeted specifically at those abruptly tasked with responsibility for cataloguing and metadata services, it also functions as an approachable introduction for any library or information worker, including students and trainees. At a slim 167 pages, the book does not pretend to provide comprehensive coverage of all cataloguing and metadata issues (such a work would surely be the most diachronic…), yet it achieves its aim of being a concise overview.

One of the most valuable aspects of the book is its capacity for brief, high-level overviews, including simple statements of summarised information which allow associations and connections to be made. For example, the authors lay out the evolution of our old friend "the catalog" from the systems developed in the mid to late 1990s (examples are listed along with their suppliers: "Innovative's Millennium, Ex Libris's Voyager, and SirsiDynix's Symphony") to contemporary products (Alma, WMS, BLUEcloud, Sierra, amongst others), alongside the development of Discovery Layers (Primo, Summon, etc.). It seems a simple thing to combine these lists of names with a brief history and explanation of their development, but it provides valuable context and clarity to those of us who have encountered these systems at various points but lacked the overview to piece together the comparable ones and match supplier names with their products.

Another stand-out few pages of the book are found at the end: "Common Acronyms Used By Cataloging and Metadata Librarians". It is often observed that the library and information world is awash with acronyms (including recursive ones – I won't forget my feeling of betrayal on discovering that the S in SPARQL stands for SPARQL) and so the value of providing a reference resource for decoding these is not to be underestimated. It exemplifies the plain-talking which this book does so well. Chapter Four opens: "While "data about data" is catchy, that definition doesn't really help us understand what metadata is all about", going on to give examples of everyday encounters with metadata on Spotify and YouTube. This helps normalise the concept of metadata and demonstrate its ubiquity.

Despite its brevity, the book manages to introduce and signpost a variety of larger issues, such as ethics in cataloguing. The treatment of this issue only covers a double-page spread, yet manages to draw brief attention to: a need to abandon assumptions of library "neutrality"; the value judgements inherent in classification; the ethical issues of defining an authorised form of someone's name; and the power dynamics and inherent biases we all bring to our work. By drawing attention to these issues even within such a short volume, the authors provide a valuable invitation for readers to consider them further, and reinforce that ethical concerns are not to be separated from practical concerns.

With the understanding that 167 pages cannot be expected to provide complete clarity on all the topics briefly covered, this book provides many valuable introductions, comprehensive lists of reference materials, definitions, lists of terms, and signposts towards further reading and resources. It need not be read cover-to-cover – in particular, chapters explaining different standards can be dipped into as needed since they are clearly labelled. Consider it a map or guidebook, rather than a manual. Chapter Five is invitingly named "Things You Might Encounter", as though you are setting off on a metadata-watching tour with binoculars in hand, and a wild periodical is about to cross your path in need of cataloguing.

I would recommend this book to students and new library workers with an interest in cataloguing and metadata, as well as the book's intended audience – managers finding themselves responsible for this area. Further, the introductory and concluding sections of the book are an accessible read for any library or information worker, and convey well the ubiquity and importance of metadata work, as well as laying out some aspects of the broader context.

This is the first title published in the new Sudden Position book series from ALCTS (putting acronym-busting into action, that's the Association for Library Collections and Technical Services, a division of the American Library Association), and it will be interesting to see what other titles are forthcoming.

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

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

Submissions: In the first instance, please contact the Co-editors:

Karen Pierce: PierceKF@Cardiff.ac.uk

Philip Keates: p.keates@kingston.ac.uk

Book reviews: Please contact the editors.

ISSN 2399-9667

CIG website: cilip.org.uk/cig

CIG blog: http://catandindexgroup.wordpress.com/

CIG twitter: @CILIPCIG

Tags from the CIG blog on specific areas of interest:

authority control book reviews Catalogue and Index cataloguing CIG activities CIGS classification committees and working groups conferences Dewey digitised material Dublin Core events folksonomies linkblog MARC metadata news RDA Semantic Web social software standards taxonomies UDC