JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

Commercial and cultural sectors: potential for data collaboration?

Graham Bell

The European Commission-funded Linked Heritage project[1] aims primarily at contributing content to Europeana, increasing the quality, richness and reuse potential of that content, and enhancing the network of expertise built up within the heritage sector by previous projects such as Athena and Minerva. But a unique facet of Linked Heritage also seeks to define how commercial organizations might engage with Europeana. This link to the world beyond libraries and other cultural memory institutions is the focus of EDItEUR[2] and its partners within the project.[3]

[1] http://www.linkedheritage.eu.

[2] EDItEUR is the trade standards body for the global book, e-book and serial supply chains. It is a not-for-profit, member-supported organisation based in London, but with a global membership of publishers, distributors, retailers, subscription agents, libraries, and system vendors. It's best known for developing the ONIX and EDItX families of metadata and transactional messaging standards, and is an acknowledged centre of excellence on metadata and identifier issues for the publishing industry. EDItEUR provides management services to the International ISBN Agency and the International ISTC Agencies, and is currently also working on projects supported by WIPO (Enabling Technologies Framework, TIGAR) and the European Commission (Linked Heritage, Arrow Plus). URL: http://www.editeur.org.

[3] Linked Heritage Work Package 4 (WP4) includes EDItEUR Ltd and the following other organizations: ICCU (Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche) – part of the Italian Min-

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5487

When discussing Europeana, of course, 'content' is actually metadata. Cultural objects, and the digital representations of those objects, remain with their host institutes. Europeana aggregates only the objects' metadata, aiming to build a comprehensive cultural discovery portal and to drive researchers, educators and students back to the websites of the originating institutes. And yet there is the 'copyright gap' – a century-long lacuna between creativity and cultural heritage. This was described in the Comité des sages's report (*The new renaissance*) as a 'black hole' of in-copyright and commercial material missing from Europe's digital cultural collections. Copyright – or doubt about copyright – can prevent the digitization of physical objects (for example, the scanning of books in libraries), and prevents institutions making digital representations of the in-copyright parts of their collections available to all via the internet. The material that cultural memory institutions deliver to Europeana is metadata describing more or less ancient objects and artefacts. Any rights and restrictions associated with the original objects, artefacts and digital representations remain in place. On this basis, the Europeana operating model is not fundamentally antithetical to commerce. However, Europeana's Data Exchange Agreement demands

istry of Cultural Heritage and Activities; mEDRA (multilingual European DOI Registration Agency Srl) – an identifier registration agency part owned by the Italian Publishers Association; MVB (Marketing- und Verlagsservice des Buchhandels GmbH) – the leading service company for the German book industry, owned by the Börsenverein des Deutschen Buchhandels, the German Publishers and Booksellers Association; NSL (National Széchényi Library) – the Hungarian National Library; Pintail Ltd – project management consultancy specializing in e-culture, library and internet technology projects; Promoter Srl– provides technical coordination and consultancy in information technology, multimedia, innovation and business development; TIB (Technische Informationsbibliothek) – the German National Library of Science and Technology. The initial report from this workgroup, written by EDItEUR's Michael Hopwood, covers metadata and identifier best practice in the commercial sector, and is available from the Linked Heritage project website, http://www.linkedheritage.eu/getFile.php?id=283.

that any rights in the metadata be waived, to allow Europeana and others to reuse and redistribute the metadata freely. Aside from the obvious difference that commercial metadata describes products that are mostly in copyright – and many of these are in commerce – some other strong contrasts need to be drawn between commercial and cultural sector metadata.

First, commercial and cultural sector metadata often describe different classes. Most cultural sector metadata is concerned with items. This is self-evident for the metadata held by an archive or a museum, as the metadata describes the individual and often unique objects or items within the collection, whether they are archaeological treasures or 19th century ephemera. For libraries, however, this is less clear: a library catalogue contains bibliographic information that is superficially similar to a national bibliography, a books-in-print database or a publisher's catalogue. But at heart, a library holdings catalogue begins as a list of the volumes in the library.[4] In familiar FRBR terms, the catalogued entities are items, with their own accession and call numbers. In contrast, a publisher's catalogue describes classes of items, or manifestations in FRBR terms, with each manifestation identified by an ISBN and comprising many individual instances or items.

Second, commercial metadata often covers a broader, richer range of data elements: a picture of the book cover, synopses of the content, extracts from the text of reviews, and a biography of the author are all common 'marketing collateral' included in ONIX for Books (ONIX is the widely-implemented standard metadata schema used in the global book trade[5]) records produced by a publisher, but not in library MARC records. There is good reason why this is so: data sells – and more data sells more. A 2011 statistical study by Nielsen (*White*

[4] But a MARC record may be more than a catalogue record, see figure 1 on page 298

[5] http://www.editeur.org/83/Overview.

Paper: The Link Between Metadata and Sales), clearly documented the positive effects of enhanced metadata on sales, whether through simple discoverability or through greater engagement with the customer. Products where a standard and very basic set of 11 metadata elements was provided saw a near-doubling of sales – both online and offline – compared with products lacking one or more of these 11 elements, and additional provision of a range of rich marketing collateral raised sales by a further 55%. There are of course other data elements required by the commercial supply chain that have no place in public-sector catalogues. The territorial nature of book rights – where a publisher may have the right to publish a work in one country but not in another – is an obvious example. This may not be familiar where a language is essentially 'national', but in English-language book publishing, it's critical for a global retailer like Amazon or Apple to know whether this product from a British publisher may also be sold in Canada or USA. There could be a different publisher or exclusive distributor who holds rights to the work in North America.

Fourth, commercial sector data is often highly dynamic. Publishers' catalogue data changes frequently. A book might be announced months before publication, and the metadata is, within that intervening period, highly provisional. Planned titles change. Publication dates change. Even author's names change. And post-publication, prices, availability, sales rights and the rich descriptive metadata are all subject to frequent updates. Commercial data is characterised by dynamic data flow rather than by static repositories of data.

Fifth, commercial sector metadata can include copyrighted content. While there is justifiable doubt over whether largely mechanical, factual bibliographic data such as title and authors could possibly be covered by copyright, publishers' metadata often includes sample text – table of contents, sample pages, perhaps even whole chapters

– that allows for no doubt. And a sui generis database right also persists over large collections of bibliographic data.

Like organizations in the cultural sector, commercial organizations commit significant resources of time and money to the creation and, more particularly, the maintenance of metadata. Maintaining rich and accurate metadata in a dynamic business environment with many thousands of new products every year is expensive – but the metadata is a key enabler for the publishing business, a core part of the process, and one that is an asset in its own right. And – somewhat ironically in the light of the growth of the open data movement – the value of that asset is growing rapidly. A decade ago, publishers employed sales teams whose sole purpose was to present books to booksellers. Increasingly, metadata is the publisher's sales team. Given the above, provision of commercial sector metadata is often accompanied by a requirement for some measure of control over the nature and context of any use made of the metadata. ONIX metadata for example often includes elements intended only for internal use within retail organizations, or data that may only be revealed publicly after some embargo date. Many publishers explicitly license use of their metadata to data aggregators or retailers, and impose restrictions on use and service level agreements on those making use of it. This might include commitments over presentation of the metadata, over accuracy and timeliness of metadata updates, over the right to redistribute the data, and above all, over clarity of business process[6] Even for those publishers that provide product metadata but forego explicit licences, an "implied licence" accompanies any metadata, and it can be argued that this limits use of the metadata to trading in, merchandising, promoting and selling the products described, and precludes redistribution. There

[6] For an example, http://www.bic.org.uk/files/pdfs/110721recipients%20best%20practice%20final.pdf.

are of course strong prima facie arguments for open licensing of data where creation of that data was publicly funded, but these do not apply where the data is created by commercial organizations. For the above reasons, commercial publishers – and organizations in other creative sectors – view product metadata as having a commercial value and sensitivity, and waiving rights to this business-critical asset would require extraordinary justification. The alternative is to strip down the range and richness of the metadata to an anodyne – and valueless – minimum, which would meet neither publishers' nor Europeana's needs. It is this issue – in effect, the construction of a business case for release of a commercially-valuable asset where all rights to that asset are waived – that will be the focus of EDItEUR's and its work package partners' effort in the second half of the Linked Heritage project.

Beyond the Europeana context, and aside from the contrasts drawn above, commercial and cultural sector metadata are in many ways complementary. In the face of budgetary pressure, many cultural sector organizations operate at least partly commercially, and publishers have long dealt with memory organizations such as libraries. There is a strong history of using commercial sector data to seed or to enrich cultural sector data.

One well-established example is the use of publishers' product metadata – in the form of ONIX records – to create CIP or MARC records for the library world. In the USA, OCLC has taken pre-publication ONIX data from publishers to construct the basis of its bibliographic records. A small British company, BDS, does the same, as part of its creation of CIP data for the British Library. The aim of these efforts is to create library-grade bibliographic records from the product records that publishers create for quite different purposes.

Carol Jean Godby of OCLC describes the process of mapping from ONIX to MARC21 records in detail, in two papers (*Mapping ONIX* *to MARC* for the latest version of ONIX 3.0; *A Crosswalk from ONIX Version 3.0 for Books to MARC 21* for the earlier ONIX 2.1). These papers present detailed 'recipes' for mapping that assert, for example, the equivalence of the ONIX <ImprintName> XML element with MARC field 260 $b or <ContributorRole> with 100/700 $e, and provide equivalent values for terms in controlled vocabularies used within ONIX and MARC.

However, such mappings are not purely syntactic, and must be constructed carefully, to ensure the maximum semantic value is carried from one record to another, without imbuing a particular metadata element with unjustified meaning and in effect 'inventing' information where nothing is implied. The two metadata schemas, and the abstract data models on which they are based, have different underlying purposes, and are not simply different ways of expressing the same information. Given the similarity of their domains, the level of semantic interoperability between ONIX and MARC is inevitably high, but not every concept in ONIX can be carried across, as many are purely supply chain-related and have no relevance to librarians or library users. Conversely, as figure 1 illustrates, ONIX for Books is not a superset of MARC – it describes only manifestations, and specifically, manifestations that are products.[7] Although an ONIX record can contain identifiers for works (FRBR expressions), this is limited to the extent that it facilitates rights trading and retail customer service.[8]

[7] In the FRBR model, books in libraries are individual items, but marc records often deal with classes of identical items (manifestations) or classes of manifestations with essentially identical content (expressions). the <indecs> model on which onix is based is similar, except that frbr expressions are indecs works. a frbr work is in effect a class of <indecs> works related to each other through revision, adaptation, translation, compilation and so on, but <indecs> models this as an inter-related group of peers rather than as a group descended from a higher-level and entirely abstract entity

[8] There is a separate ONIX metadata format used to characterise <indecs> works –

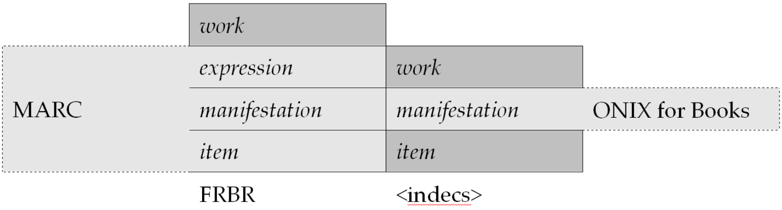

Figure 1: Rough equivalence of MARC and ONIX entities.

Despite these caveats, as Godby writes,

> the outcome of the [mapping] is a MARC 21 record with AACR2 semantics that can be automatically generated from an ONIX 3.0 source, pass a rigorous semantic validation, serve as a rough draft that can be further refined by cataloging best-practices guidelines, and qualify for inclusion in a quality-controlled library database.

Of course, what results from such a mapping is not always a library-grade record, as libraries remain more concerned than publishers with – for example – authority files and cataloguing rules, and ONIX records are not always complete because few data elements are mandatory. But the process of mapping is effective, efficient, and means that cataloguing processes can begin long before the book is available. Mapping from ONIX to MARC21 illustrates how commercial metadata can seed and enrich cultural sector metadata. But interoperability is two-way: cultural data can in principle be used to enrich existing commercial data too. The potential for this can be seen in the new International Standard Name Identifier (ISNI) for public identities of parties involved in creative endeavours.[9]

ONIX for ISTC Registration, used for the registration of ISTC (International Standard Text Code) work identifiers.

[9] http://www.isni.org.

The standard has been launched with around a million identities pre-defined, based on data from national library authority files, and use of the ISNI enables commercial metadata to differentiate between, say Prof. Richard Holmes (ISNI 0000 0001 2147 5396) and the identically-named Prof. Richard Holmes (ISNI 0000 0001 1768 5542), or to state authoritatively that Julian Cope the musician is the same party as Julian Cope the author (ISNI 0000 0000 7725 4712). Linked Heritage's predecessor project Athena[10] delivered a data mapping engine called MINT (Metadata Interoperability Services), a data schema LIDO (Lightweight Information Describing Objects), and a LIDO to Europeana (ESE) mapping. The current focus of EDItEUR and its project partners within the Linked Heritage project is on building mappings within MINT that are conceptually similar to the ONIX to MARC work outlined above. This will enable large volumes of ONIX metadata – and commercial data from other creative sectors, including recorded music (DDEX metadata), film and TV (EIDR metadata), and photography (IPTC metadata) – to be mapped into LIDO, and potentially delivered (either in whole or in part) into Europeana. The appeal of an enriched Europeana record for a van Gogh painting – say The Café Terrace on the Place du Forum, Arles, at Night – with links to the latest commercial biography of van Gogh, a modern travel guide to the city of Arles, a commercial recording of César Franck's Symphony in D minor (completed only a few days before the painting), and perhaps a contemporary photo from a picture library of the café terrace on the Place du Forum, is clear.

Listing 1: Equivalent ONIX and RDF metadata expressions.

```
<Contributor>
  <ContributorRole>A01</ContributorRole>
  <NameIdentifier>
```

[10] http://www.athenaeurope.org.

```
  <NameIDType>16</NameIDType>
  <IDValue>0000000121479135</IDValue>
 </NameIdentifier>
 <PersonNameInverted>Sjöwall, Maj</PersonNameInverted>
</Contributor>

   http://ns.editeur.org/onix/3.0/reference/Contributor
       genid:A96837
genid: A96837 http://ns.editeur.org/onix/3.0/reference/
    ContributorRole  http://ns.editeur.org/onix/ codelists
    /17#A01
genid: A96837 http://ns.editeur.org/onix/3.0/reference/
    NameIdentifier  "0000000121479135" of type http://ns.
    editeur.org/onix/codelist/44#16
genid:A96837 http://ns.editeur.org/onix/3.0/reference/
    PersonNameInverted "Sjöwall, Maj" of type http://ns.
    editeur.org/onix/ codelists/18#01
```

MARC21 and its associated English-speaking AACR2 cataloguing rules are destined to be replaced by RDA cataloguing and some yet-to-be-defined data format[11] – and this route is likely to followed by other flavours of MARC too. The destination of this journey is 'Linked Data' in some form, and it is this that holds the promise of automatically associating the metadata record for Vincent's painting with that for Cesar's symphony, thereby enriching both. Yet what we have now can best be described as 'data with links': ONIX metadata contains information linking books to people, to places, subjects, dates, other books, and the underlying data could be re-expressed in RDF as illustrated in listing 1.[12] There is an explicit ONIX data model (separate from the XML schema) to guide this re-expression. This

[11] http://www.loc.gov/marc/transition.

[12] The four RDF triples use an arbitrary blank node (a96837) to represent the contributor, and the node has three properties representing the role, name and identifier of the contributor. the use of URIs in the RDF syntax is a more easily machine-

type of adaptation is conceptually similar to mapping between ONIX and MARC, though the first step to mint and promote the canonical URIs necessary for expressing the ONIX as linked data has not yet been taken. The benefit of re-expressing ONIX (or other commercial metadata) as Linked data is that it may be simpler to process the links expressed within the data automatically. But ultimately, this may not be enough. Linked data using industry-specific vocabularies and proprietary identifiers tends to form islands of data, richly linked internally, but ultimately not well linked to the rest of the Linked Data cloud. To increase the density of links between these islands of data, it's necessary to add a semantic mapping layer that says – in effect – this term for a relationship or RDF predicate used in this industry sector is the same as that term used in a different sector. Listing 2 shows how such semantic mappings can be expressed.

Listing 2: Sample RDF showing semantic relationship between onix contributor role (*a01*, meaning 'written by', used in the second triple in figure 2), the exactly equivalent marc relator *aut* and the broadly equivalent *authorwork* term from RDA.

```
<skos:Concept rdf:about="http://ns.editeur.org/onix/codelists
    /17#A01">
  <skos:inScheme rdf:resource="http://ns.editeur.org/onix/
      codelists/17#"/>
  <skos:notation rdf:datatype="http://www.w3.org/2001/
      XMLSchema#token">
    A01</skos:notation>
  <skos:prefLabel xml:lang="en">Written by</skos:prefLabel>
  <skos:exactMatch rdf:resource="http://id.loc.gov/vocabulary/
      relators/
    aut"/>
```

processable variation on onix codelists (controlled vocabularies). Note that canonical URIs for expressing ONIX metadata in Resource Description Framework (RDF) have not been published – this is merely an illustration. the subject of the first triple is omitted, as it is in the ONIX, but could be an identifier for 'the book' such as an ISBN.

```
    <skos:closeMatch rdf:resource="http://rdvocab.info/roles/
        authorWork"/>
  </skos:Concept>
```

Similarly, some agreement on public identifiers used for common entities – people and their public identities, places, organizations etc, is necessary. If each heritage and commercial sector uses a different sector-specific or proprietary identifier for a public identity, for example, then making links between sectors becomes reliant on the error-prone process of matching names. The use of a common, cross-sector public identifier – ISNI in this case – solves this issue.

So when large volumes of data from a range of commercial and cultural sectors are aggregated, the interconnectedness of the data – the degree to which data from one sector can enrich that of another – is dependent on careful semantic mapping and the use of identifiers rather than textual names. It is the use of common public identifiers, interoperable semantics and shared vocabularies that is the glue that allows triples to be bound together automatically, inferences made and implicit connections to be revealed. Without these, disparate databases cannot be bound into a single data space.

References

Breedt, Andre and David Walter. *White Paper: The Link Between Metadata and Sales*. Woking: Nielsen Book, 2011. http://www.isbn.nielsenbook.co.uk/uploads/3971_Nielsen_Metadata_white_paper_A4(3).pdf. (Cit. on p. 293).

Godby, Carol Jean. *A Crosswalk from ONIX Version 3.0 for Books to MARC 21*. Dublin, Ohio: OCLC Research, 2012. http://www.oclc.org/resources/research/publications/library/2012/2012-04.pdf. (Cit. on p. 297).

—–. *Mapping ONIX to MARC*. Dublin, Ohio: OCLC Research, 2010. http://www.oclc.org/resources/research/publications/library/2010/2010-14.pdf. (Cit. on p. 296).

The new renaissance: Report of the "comité des sages" on bringing Europe's cultural heritage online. Bruxelles: European Commission, 2011. DOI: 10.2759/45571. (Cit. on p. 292).

GRAHAM BELL, EDItEUR.
info@editeur.org

Bell, G. "Commercial and cultural sectors: potential for data collaboration?". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5487. DOI: 10.4403/jlis.it-5487. Web.

ABSTRACT: The main goals of the Linked Heritage project (sponsored by Europeana) are to provide qualified content to Europeana from the public and private sector. To this aim is created WP4 (Work Group 4) in which the organization EDItEUR takes part. The 'content' are metadata for the cultural heritage. Starting by addressing the issue of the "copyright gap", which can involve metadata, the article notes the differences between metadata types developed by the private sector (ONIX for books) and those defined by the public one (FRBR, MARC, MARC21). The aim is to develop integration of both sectors into a shared references core. Exploring the common reference framework requirement, the article emphasizes the new International Standard Name Identifier (ISNI) potential, which allows to uniquely identify the subjects involved in the creative field. The future outlook can be further enhanced by involving additional metadata mapping that relates books, people, places, data, other books and other references in a possible 'Linked Data' form, within which priority should be given in common public identifiers, related semantic mapping layers and shared vocabularies.

KEYWORDS: Europeana; ISNI; Library linked data; Linked Heritage Project; EDItEUR

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15