Explorations in Linked Data practice for early music corpora

Tim Crawford, Ben Fields, David Lewis
Department of Computing
Goldsmiths, University of London
United Kingdom
{t.crawford, b.fields, d.lewis}@gold.ac.uk

Kevin Page
Oxford eResearch Centre
University of Oxford
United Kingdom
kevin.page@oerc.ox.ac.uk

ABSTRACT

Exploring connections between pieces, people and places and relating them to culture as a whole is a central activity of musicology. As libraries increase the availability of musical information in digital form, the data available for such research also expands, but to take such resources together and combine them with others that are relevant a further step of alignment and linkage is needed. We describe here the process and tools we applied to two corpora of early modern music: Early Music Online, which comprises catalogue metadata in MarcXML and facsimile images for approximately 8,500 items of early printed music; and the Electronic Corpus of Lute Music, containing over 1,000 pieces with supporting metadata. A supervised process with automated elements assists the musicologist to create a linked and extensible knowledge structure, aligning entities within and between corpora and to external Linked Data. Finally, we reflect upon how we believe these methods integrate with, and indeed form a crucial element of, the transformed process of modern digital scholarship.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data sharing*

General Terms

Data modeling, Data alignment, Digital humanities

Keywords

Linked Data, Semantic Web, Musicology

1. INTRODUCTION

Music of the 16th and 17th centuries has reference and quotation in its nature, and is the product of a culture of connection, copying and redistribution. As printing techniques improved music publishing became an increasingly viable business. Musicologists use details of book printing, contents and transmission to help understand many aspects of 16th-century musical culture, such as instrumental music[2], and the spread of 'soft power' through colonial and other political relationships beyond Europe [3] (a more general discussion of printed music books and culture can be found in [4], amongst others). Such connections of pieces, people and places form networks; exploring these traditionally uses limited assistance from online resources. The use of authority lists in library catalogues helps for internal navigation, and for consistency, but becomes much more useful when external links are added, either to other collections or to wider sources of cultural, historical and geographical information.

In this paper, we describe an approach to structuring, aligning and linking entities within pre-existing catalogues, as a first step towards supporting humanistic research. We first describe the corpora we use, before considering the structure of the catalogue information and how the data was transformed for our purposes; we then discuss Linked Data, the process of alignment and the publishing of RDF data. Finally, we consider future work and the benefits that will be delivered to scholarship as a result.

2. CORPORA

The two collections of early music that we selected overlap in scope and musical content, but were catalogued in very different ways and for different purposes, making the task of modelling and aligning the information they contain challenging, but a useful and realistic test of our approach.

Early Music Online (EMO)[1] presents digitised images of 324 16th-century music books from the British Library (eg. figure 1), and contains about 8,500 pieces of music. Metadata, expanded from the original BL records, was provided to us as MARCxml. Mostly vocal music in part books, with each voice printed in a separate volume, EMO also includes music for keyboard instruments and for lute, including many arrangements of vocal items which can also be found in their original form within EMO.

The Electronic Corpus of Lute Music (ECOLM)[2] has the overall goal of providing a machine-readable corpus of European lute music in order to make this historically important repertory accessible to a wider audience. The current ECOLM MySQL database includes metadata for about 2,463 pieces, over 1,200 of which are also present as full-text

[1] www.earlymusiconline.org
[2] www.ecolm.org

Digital Libraries 2014 London, UK

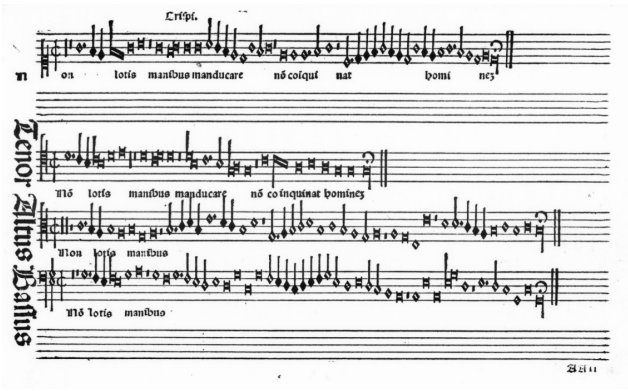

Figure 1: The first page of music from Petrucci's *Motetti de passione, de cruce, de sacramento, de beata virgine et huiusmodi* (Venice, 1503) from Early Music Online

encodings of the music, mostly from the 16th century. From the lute books in EMO, between 300 and 1,000 more encodings have been added to ECOLM using Optical Music Recognition (OMR).

These two digital library resources have significant chronological and repertorial overlap and are of potentially great musicological value; the aim of the work here is to provide the basic layer of identification and linking of concepts associated with the resources, making them more accessible and reusable for future scholars.

3. SOURCE DATABASE STRUCTURES

MARCxml is a serialisation of standard MARC records, permitting a wide range of granularity, level of detail and cataloguing policy for any given catalogue. The EMO metadata shows some variation in quality and style. Where authority lists are applicable, an internally generated ID hash is given, uniquely identifying the entity, such as a person or place, but without explicit reference to any external provider. Since the records are designed to be self sufficient, string descriptors are also given in these cases, and are not always identical between references to the same entity.

Since BL cataloguers were concerned primarily with locating the physical item rather than individual parts, the titles and attributions of musical contents of a source are listed in a single, often repeated, data field (`505 - Formatted Contents Note`).[3] Works are separated by a double dash, `--`, and attributions by a slash, whilst subdivided works are usually – but not universally – indicated by brackets and the Latin for the part number (see figure 2). These attributions are, by policy, given as in the source. Associations with names from authority files are in a separate data field (`700 - Added Entry-Personal Name`), which provides a list of names (and identifiers) but cannot directly connect them with individual inventory items and does not routinely specify the roles that those individuals played in the history of the source and its contents.

[3]Another reason for this is the inability of MARC to support deep tree structures, at least below the subfield level. Placing so much information in strings usually designed to be human-readable also limits the ability to insert unique identifiers at these deeper levels.

```
<marc:datafield tag="505" ind1="0" ind2="␣">
  <marc:subfield code="a">Non lotis manibus / Crispi. -- O
       D[omi]ne Jesu [Christe] adorote in cruce (In
      contents: Officium de passione) (secunda pars: O d[
      omi]ne Jesu [Christe] adorote in cruce; tertia
      pars: O domine Jesu christe adorote in sepulcro;
      quarta pars: O domine Jesu christe pastor bone;
      quinta pars: O domine Jesu christe p[ro]p[ter]
      illam) / Josquin -- Qui velat[us] facie fuisti (
      secunda pars: Hora [que] duct[us]; tertia pars: In
      flagelis potu[m] fellis; quarta pars: Honor [et] b[
      en]edictio; quinta pars: In amara crucis; sexta
      pars: Qui jacuisti mortuus; septima pars: Christu[m
      ] duce[m] redemit nos) / Josquin -- Secundu[m]
      multitudine[m] / [Anon] -- [\ldots{}] -- Ave domina
       s[an]cta maria / [Anon] -- Parce d[omi]ne parce
      populo tuo / Franci --</marc:subfield>
</marc:datafield>
<marc:datafield tag="505" ind1="0" ind2="␣">
  <marc:subfield code="a">Lauda syon salvatorem (secunda
      pars: In hac me[n]sa novi regis; tertia pars: Sub
      diversis specieb[us]; quarta pars: A sume[n]te no[n
      ] [con]cisus; quinta pars: Sumu[n]t boni sumu[n]t
      mali; sexta pars: Fraco demu[m] sacrum; septima
      pars: Ecce panis angelor[um]; octava pars: Bone
      pastor panis vere) / Brumel -- [\ldots{}]-- Sic
      unda impellitur unda / [Pierre Moulu].</
      marc:subfield>
</marc:datafield>


<marc:datafield tag="700" ind1="1" ind2="␣">
  <marc:subfield code="a">Petrucci, Ottaviano</
      marc:subfield>
  <marc:subfield code="0">c378d380855a012e8330fd1cbfbf31ec
      </marc:subfield>
</marc:datafield>
<marc:datafield tag="700" ind1="1" ind2="␣">
  <marc:subfield code="a">Stappen, Crispin van</
      marc:subfield>
  <marc:subfield code="d">ca. 1465-1532</marc:subfield>
  <marc:subfield code="0">c9dc9000855a012e8330fd1cbfbf31ec
      </marc:subfield>
</marc:datafield>
<marc:datafield tag="700" ind1="1" ind2="␣">
  <marc:subfield code="a">Josquin</marc:subfield>
  <marc:subfield code="d">d. 1521</marc:subfield>
  <marc:subfield code="0">600ca700855a012e8330fd1cbfbf31ec
      </marc:subfield>
</marc:datafield>
```

Figure 2: The contents (above) and first three attribution entries (below) for Petrucci (1603)

Since all the EMO sources are prints rather than manuscripts, their place of origin – or at least of printing – is usually clear and unambiguous. In the MARC, this is indicated by specifying city and country (data field `752 - Added Entry-Hierarchical place name`).

4. LINKED DATA AND THE ALIGNMENT PROCESS

As well as reconciling entities and their relationships between the two source datasets we wished to enable the scholar to situate the early music resources within the context provided by other external repositories of knowledge. Similarly, as scholars work with data from other sources and create new knowledge we want to enable them to reference accurately and unambiguously or incorporate data we publish from EMO/ECOLM – the overall effect being a growing multi-linked web of knowledge. For this we employ semantic web technologies, minting URIs for the entities identified in the combined database and creating linked data between relevant external and internal resources, reusing existing ontologies where possible.

The work described here is not intended as a completed state of all possible links from the source resources, but a foundation upon which scholars can continue to extend, supplement, and potentially replace alignments of varying degrees of certainty – since the act of alignment is one of scholarship and specialist knowledge in of itself, as we will see below. The technologies applied here are particularly suited

		Links		
		Total	VIAF (%)	DBPedia (%)
Settlements	EMO	24	24 (100%)	24 (100%)
	ECOLM	124	105 (85%)	111 (90%)
	Total	148	129 (87%)	135 (91%)
Countries	EMO	7	7 (100%)	7 (100%)
	ECOLM	16	16 (100%)	16 (100%)
	Total	23	23 (100%)	23 (100%)

Table 1: Most settlements and all their containing countries from this dataset could be aligned successfully with VIAF and DBPedia.

to the tasks, due to their support of "open ended" knowledge: there may be other parts of the knowledge structure distributed elsewhere on the web, and any specific resource can have assertions stated about it that can conflict with previous ones (although the scholarly record may prefer explanation). In this we are mindful of previous work regarding "layered" approaches using linked data ([1]), a technique that can be combined with datasets such as ours.

Since the number of fields in the dataset that could be linked is large, we selected a subset to form our initial published information. These features should be as useful as possible by themselves, they should link in meaningfully with external resources, they should be relatively simple to model and factually be relatively uncontroversial. Ideally, the data aligned in the first stage should also be useful for limiting the search space for subsequent alignment efforts.

For these reasons, we chose to focus alignment efforts on people and places, particularly places of printing or publication, which are usually clearly recorded on the documents concerned, and which provide much important context for studying the culture surrounding them.

Since links to external authorities would also act as internal links between our two collections, we focussed our alignment efforts on external Linked Data sources that offer potential for further useful linking. The EMO catalogue is based heavily on BL data, which currently uses Library of Congress authority lists to standardise entries. This suggested that one set of external links should be to these authority lists, or rather to VIAF, the Virtual International Authority File (viaf.org), which aligns the various library authorities into a single resource.

VIAF is useful for information published as linked data by other libraries, but the number of outward connections beyond the library domain is variable. For this reason, two other Linked Data resources were also used: DBPedia, the semantic-web interface to Wikipedia, which was used for place names; and MusicBrainz, an open online and rapidly growing database for recorded music, which we have used for personal names.

The number of distinct geographical places in EMO and ECOLM is quite limited – 24 and 124 respectively – and relatively uncontroversial (in identity if not always in reliability of association), so alignment to VIAF and DBPedia could be carried out manually. The results of these matches are shown in table 1. Countries, and cities with printing presses in the 16th century, are sufficiently important entities that any geographical authority would be expected to have full coverage. Place names in ECOLM however, unlike those in EMO, may also relate to individuals involved in the music or its sources, to places of publication of secondary literature or to locations associated with manuscripts, all of which make

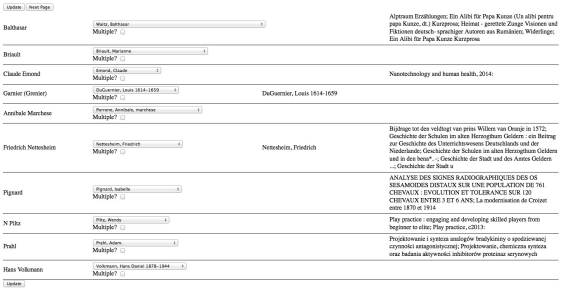

Figure 3: Screenshot from the web application used for aligning personal names with external resources

alignment trickier, and the probability of suitable URIs existing in a given authority list significantly lower.

It is important to note that, although settlements and countries in a given list may be uncontroversially identifiable, this is not always the case. Furthermore, the containing political unit of any given settlement may be fluid – the containing country, for example, may be historically variable – and so the 'correct' country to indicate will depend on the user need or the exact predicate used. Our policy is to ensure that the smaller geographical unit is adequately identified with an external URI that can be annotated later with more detailed historical information and, at the same time, to indicate the modern containing country for the purposes of convenient retrieval.

Although place names can be ambiguous, they seldom are so in the collections considered here, since most relate to large centres of commercial or musical activity. Duplicate place names are relatively rare in ECOLM and EMO, and manual alignment, on a small scale, can be done quickly. For personal names, this is not so simple. Names may be ambiguously specified, perhaps as a family name or an honorific only.

Names also frequently recur within the same time and culture, with sons taking on the family name as well as trade. Furthermore, since there exists no single resource of composers, URIs and complete (attributed) work-lists – VIAF entries may even have no works named at all – it may be difficult to ensure that entries in authority lists refer to the person specified in the source or catalogue entry.

Since the number of people to be aligned is an order of magnitude larger than the number of places, a manual approach was not considered feasible. Instead, string searches (including date strings, where available) were carried out through the VIAF SRU API and MusicBrainz API. For the 1,235 persons identified by either the EMO or the ECOLM database, 5,653 matches were retrieved from VIAF and 1,103 from MusicBrainz.

A purpose-designed web application was developed which presents pages of candidate matches, including work titles and alternative names as provided by the authority source (figure 3). The user then chooses one or more matches or 'unknown' to indicate that either the list provided didn't include anything relevant or that there was insufficient data presented to inform a choice. Table 2 summarises the result of this supervised process.

The lower proportions in this table reflect both the ambiguity of the data and the difficulties of resolving it quickly. Many people named in either resource weren't directly in-

volved with the music, such as dedicatees or printers, or authors of secondary literature. Such people rarely occur in MusicBrainz, explaining the threefold difference in matches between the two authorities.

			Links	
		Total	VIAF (%)	MusicBrainz (%)
Persons	EMO	1,037	661 (64%)	243 (23%)
	ECOLM	198	150 (76%)	57 (29%)
	Total	1,235	811 (66%)	300 (24%)

Table 2: Alignment of personal names is less successful than places, but significant numbers *do* align.

Of the verified matches, VIAF's 811 results contain 778 distinct URIs, whilst the 300 results from MusicBrainz have 272. These duplicates were partly due to overlap between the two databases, but also revealed instances in the EMO metadata where the same person, with a slightly altered name or date string, had received several distinct identifiers.

These confirmed alignments are especially useful if they facilitate the discovery of more linked information. A simple search using sameas.org, for example, connects 194 of the discovered VIAF person URIs to DBPedia, 120 to DBtune, 110 to the BBC and many others to individual library catalogues. 277 of the entries identified in VIAF were also identified in MusicBrainz – information currently represented using the Similarity Ontology, but which could be published as novel sameAs connnections.

5. PUBLISHING THE RDF

We presumed that predicates and ontologies should be selected pragmatically, expecting that small adjustments might be necessary later, but that the initial selections should be robust. Where possible, ontologies and predicates should be drawn from domain-relevant sources likely to be used by others dealing with similar materials. Avoiding the generation of our own concepts was also a priority. Table 3 shows the predicates and ontologies we used.

Concept	Predicate	Ontology
Settlement	`dbp:Settlement`	DBPedia
Country	`dbp:Country`	DBPedia
Book	`bibo:Book`	Bibliographic ontology
Person (creative role)	`foaf:Maker`	Friend of a Friend
Publication	`blt:Publication`	British Library
Publication year	`event:Time`	Event ontology
Publication place	`event:Place`	Event ontology
Alignment	`sim:Similarity`	Similarity ontology

Table 3: Some of the main concepts and predicates used and their sources.

The database of combined information from EMO and ECOLM has been published as RDF using a `d2r` instance, mounted at `http://slickmem.data.t-mus.org`; this also provides a browsable HTML view of the published data.

6. FUTURE WORK

Working at the level of the book rather than the musical work limits the usefulness of the EMO resource for musicological research. To address this, we are developing web software to speed up the process of correcting inventory lists and indicating where on images a given work is set.

Identifying works on the basis of textual titles and attributions is difficult, especially in a period where titles of instrumental arrangements are often strangely mis-spelt. Titles using the opening lyric can be confusing in cases where the same poem or prayer is set to music more than once, and completely unhelpful for common parts of the liturgy such as the Ordinary of the Mass. Furthermore, instances of the same conceptual work are seldom identical, and their relationships must also be approached with caution. Once the results of OMR on EMO books are associated with items in the inventory lists, Music Information Retrieval techniques can be used to help identify related items, greatly enriching this research infrastructure.

7. DISCUSSION

We have described the importance of connections, whether in terms of content or of context and history, in the study of musical documents. These links are commonly identified by librarians and scholars and may be recorded directly in catalogues, or in separately published scholarship. In the former case, these are still often only internal connections within a library, whilst in the latter case, they are rarely in machine-readable form. Linked Data gives the possibility to provide machine-readable, outward facing connections that can be annotated and expanded iteratively and collaboratively. Each enrichment and alignment stage adds functionality, but also makes subsequent alignment tasks easier.

Most of our alignment work with these collections requires some manual labour. Assistive alignment, using software which suggests credible options and supports the manual process, will make the task both faster and more reliable. Most importantly, publishing these connections as Linked Data provides for reuse, amendment and, sometimes inevitably, even disagreement. By publishing our alignment information as judgements of similarity, together with the workflow that generated those judgements, we hope to allow for evaluation, discussion and scrutiny and thus to accommodate at least some of the the academic discourse that is often absent from resources that are often presented as purely 'factual'.

7.1 Acknowledgments

The authors are supported by the AHRC-funded *Transforming Musicology*; research reported here was funded as a mini project by the EPSRC *Semantic Media Network*. *EMO* was funded by JISC, and *ECOLM* by the AHRB and AHRC.

8. REFERENCES

[1] S. Bechhofer, K. Page, and D. De Roure. Hello Cleveland! linked Data publication of live music archives. In *14th International Workshop on Image and Audio Analysis for Multimedia Interactive*. IEEE, July 2013.

[2] J. Griffiths. Hidalgo, merchant, poet, priest: the vihuela in the urban soundscape. *Early Music*, 37(3):355–366, 2009.

[3] D. R. M. Irving. The dissemination and use of European music books in early modern Asia. *Early Music History*, 28:39–59, 2009.

[4] M. S. Lewis. The printed music book in context: Observations on some sixteenth-century editions. *Notes*, 46(4):899–918, June 1990.