A linked research network that is *Transforming Musicology*

Terhi Nurmikko-Fuller and Kevin R. Page

Oxford e-Research Centre, University of Oxford, United Kingdom
{terhi.nurmikko-fuller,kevin.page}@oerc.ox.ac.uk

Abstract. Semantic Web technologies offer solutions for bridging discrete and even disparate datasets. Linked Data has been seen in several Digital Humanities projects, but through the alignment of instance-level entities rather than the capture of workflows, which have yet to become part of the publication paradigm for reporting on completed research. In this paper, we assess the functional requirements of digital Musicology research questions, and propose ways for using the inherent semantics of workflow descriptions alongside instance data to link them. We report on the design of a linked research network for Musicology.

Keywords: Musicology, linked research network, semantic requirements

1 Introduction

Collaborative scholarship brings together academics, diverse datasets, and different research foci. An example of this is *Transforming Musicology*,[1] an exploration into the ways digital technologies can influence the future development of scholarship on music, whether it is represented as sound, score, or symbol. This interdisciplinary endeavour bridges projects from 14 Universities, all with idiosyncratic methodologies, workflows, research agendas, and data. We report on the iterative process of assessing the needs and requirements of an underlying linked research network, which uses Semantic Web technologies to connect these projects by drawing in elements from different sources, resulting in a complementary combination of resources for the scholars involved, and beyond.

This use of Semantic Web technologies to capture workflow is not without precedent [12], but whilst the value of reproducible investigative processes has been noted in Natural Sciences and Bioinformatics [9], it has yet to be adopted as the norm in the publication of research in the Digital Humanities. Using workflow metadata as the semantic glue within the linked research network helps by-pass the "knowledge burying" problem described by Mons [13], who critiques the prevalent practice of publishing final analysed datasets only. The importance of workflow capture for the purposes of reproducibility in the sciences has been noted by Bechhofer, *et al.* [1], and the benefits of doing so extend to the reuse

[1] http://transforming-musicology.org/

of processes developed for one project in the context of another (e.g. to alleviate labour intensivity).

In *Transforming Musicology*, we enrich instance level data connections (see Section 2) with the semantics of workflows. The methodologies of each constitutent project were recorded and systematically assessed for opportunities of support and reuse. Workflows were divided into four consecutive, tripartite steps: *data preparation*, *data capture*, *summarizing*, and *visualization*. Each has input data, a process, and resulting output. Metadata semantics capture the relationships, provenance, and other aspects of each part of the workflow, including dependencies and causation (e.g. prov:wasDerivedFrom from Prov-O [19]).

There are eight ***areas of study*** (AS). The core (AS1 – 3) are under development by the Universities of Oxford and London, Goldsmith's College – these are supplemented by investigations at other institutions (AS4 – 7):
AS1: 16th century lute and vocal music that combines tablature with audio [6];
AS2a: Analysis of leitmotivs within the compositions of Richard Wagner [18];
AS2b: The psychological effects these leitmotivs can have on the listener [14];
AS3: Social media of Musicology, concentrating on Genius[2] and Echonest;[3] and
AS4: *Medieval Music, Big Data and the Research Blend* (Southampton) [5];
AS5: *Characterising stylistic interpretations through automated detection of ornamentation in Irish traditional music recordings* (Birmingham; Birmingham City; and the Dundalk Institute of Technology)[10]; the other multi-institutional
AS6:*In Concert: Towards a Collaborative Digital Archive of Musical Ephemera* (Cardiff; Birmingham; British Library; Goldsmiths College; and Illinois) [7]; and
AS7:*Large-scale corpus analysis of historical electronic music using MIR tools: Informing an ontology of electronic music and cross-validating content-based methods* (Durham).

2 Semantic Overlap

(AS3), (AS5), and (AS7) overlap in the temporal scope of the datasets; (AS4) is an isolate. (AS6) can bridge (AS1) with (AS2) (see Figure 1). They share data types such as .csv and .jpeg; (AS1), (AS2a), (AS3), (AS4), and (AS6) all analyse text and content, whilst (AS1), (AS5), and (AS7) contain an audio component. (AS2), (AS3), and (AS6) contain known instances of shared entity-level data. All but (AS3) and (AS4) largely focus on resource metadata at the *data capture* stage of the workflow.

Methodological parallels are limited to similar tools, e.g. (AS5) uses Sonic Visualiser,[4] (AS1) utilises Sonic Annotator.[5] The extent to which automated process are relied on varies from one (AS) to another – they are most actively used in (AS5). (AS6) has exports in JSON; (AS1) in XML. (AS4) data is stored

[2] http://genius.com/
[3] http://the.echonest.com/
[4] http://www.sonicvisualiser.org/
[5] http://www.vamp-plugins.org/sonic-annotator/

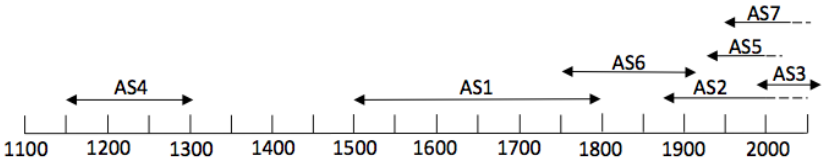

Fig. 1. Temporal overlap of areas of study (AS) in *Transforming Musicology*

in an instantiation of ePrints,[6] and metadata can be exported in a number of different formats, including JSON, XML, and RDF (mapped to a custom ontology). The projects make use of a range of existing repositories (e.g. ePrints), flat files, spreadsheets, and relational databases (MySQL). Whilst the shared aim is to publish Linked Open Data (LOD), the necessary mapping and data conversion methods differ.

3 Illustrative Musicological Research Questions

Following Bechhofer, et al.[1], we produced five hypothetical scenarios for illustrative purposes to describe possible ***research questions*** (RQ). These arise and encompass elements from more than one (AS):
RQ1: Alice discovers Bob used the NNLS Chroma plug-in[7] for Sonic Annotator to extract features from 16th century lute music. She needs access to Bob's dataset to verify his results, and to the tool to repeat the workflow on her data.
RQ2: Casey studies the publication paradigms and prosopography of printers in the 16th century: are there patterns, hubs of activity, and genre-specializations?
RQ3: David finds lyrics sung by *Siegfried* (a character in Richard Wagner's *Der Ring des Nibelungen*) on Genius. He needs complementary information (text companions, audio, notations, images) to establish an interpretative framework.
RQ4: Edward is interested in communities of practice around digital Musicology. He wants to identify pioneering institutions, preeminent scholars, to find answers to frequently asked questions, and to receive guidance on best practice.
RQ5: Frankie has annotation data captured during a live operatic performance. He is looking to represent the semantics of the annotations as RDF, and merge them with existing data already in a triplestore.

The ***functional requirements*** (FR) of the (RQ) were systematically assessed through an iterative process in response to a Request for Proposal: the details of each scenario were identified, and possible solutions proposed. Off-the-shelf tools and resources are recommended where available (see Section 4). The aim was to find commonalities between the needs of the (RQ): addressing these enables the integration between disparate datasets, but also between the raw data and the user, who is free to analyse and interpret data in the context of

[6] http://eprints.soton.ac.uk/
[7] http://isophonics.net/nnls-chroma

their own research agenda. Scholars are in a position to benefit from the output of other (AS) for their analyses.

Table 1. Functional Requirements (FR) for Research Questions (RQ)

FQ	Function description	RQ1	RQ2	RQ3	RQ4	RQ5	Tooling
FR1.1	Document repository with search & upload	✓	✓	✓	✓	✓	T1
FR1.2	Code repository with search & upload	✓	✓	✓	✓	✓	T2
FR1.3	Audio repository with search & upload	✓		✓	✓	✓	T3
FR1.4	Metadata repository	✓	✓	✓	✓	✓	T4
FR1.5	Image repository with search & upload			✓	✓	✓	T5
FR1.6	SPARQL endpoint	✓	✓	✓	✓	✓	T6
FR1.7	API	✓	✓	✓	✓	✓	T7
FR1.8	Overarching ontology	✓	✓	✓	✓	✓	T8
FR1.9	NLP tools		✓				T9
FR1.10	Niche ontologies		✓		✓	✓	T10
FR1.11	Data visualization		✓			✓	T11
FR1.12	Social network analysis		✓	✓	✓		T12

In the absence of a centralised structure for the sharing and amalgamation of information, Semantic Web technologies support access to, and the exchange of, data across all areas of study. The idea of a system incorporating a number of different types of servers (image, document, audio, etc.) bridged by a data sharing platform began to form. The vision of a coherent collection of metadata for all resources, data, tools, and code, emerged.

4 Conclusion and Future Work

As illustrated by Table 1, many of the FRs outlined above can be addressed with existing, off-the-shelf ***tooling*** (T). Repositories are an example of this: ePrints (T1), where fully and semi-automated processes allow for metadata extraction as RDF; Zotero[8] (T2), a solution for the archiving and long-term storage of code and tooling with the added benefit of the establishined workflow for importing from GitHub, which is used as a development environment with version control; triplestores as metadata repostiories (T4); and ResearchSpace[9] (T6), which provides a graphical user interface to a triplestore, allowing Musicologists to query of the underlying RDF metadata without using SPARQL [15]. Although configured to use Blazegraph[10] and the CIDOC CRM [3], ResearchSpace is both triplestore

[8] https://www.zotero.org/
[9] http://www.researchspace.org
[10] https://www.blazegraph.com/

and ontology agnostic, and can be used with Virtuoso,[11] and a purpose-built ontology (T8) that incorporates classes and properties from a number of known OWL ontologies, such as (but not limited to) the Music Ontology[8], Event [17], Timeline [16], Prov-O, and Research Objects [4], and is designed to be sufficiently flexible to allow for the future integration of the structure designed as part of (AS2a). For the audio repository (T3), *Tranforming Musicology* is in a position to benefit from earlier Musicological projects [2]; for images (T5), IIIF[12]-compliancy is highly desirable, making Loris[13] (an open source, Python-based image server) the repository of choice. Known social networking analysis tools (T12) can support (AS3) and any Musicological prosopography occuring in other (RQ). Where applicable, instance level alignments to external authorities such as VIAF[14] and Musicbrainz[15] can be implemented. Visualization techniques used in (AS6) can be reapplied (T11) to support other (AS).

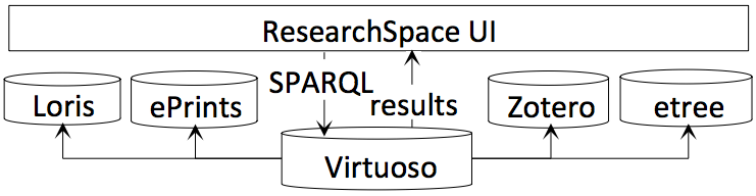

Fig. 2. An architectural realisation to address the FRs of RQ5

Some aspects of the linked research network require new development. These include identifying necessary APIs (T7) and establishing their interaction with any future graphical user-interface implementation; an over-arching ontology, as described above (T8), to connect smaller, more domain-specific models (T10); and for (RQ2), a natural language processing tool (T9), which builds on an earlier prototype by Khan et al [11].

This assesment of (FR) illustrates the large numbers of readily available existing tools, and pinpoints those circumstances where new builds are necessary. Such assesments are valuable in the planning and implementation of research projects, helping maximise potential linkage (e.g. through shared schema) and to minimise development overlap. The resulting linked research network will aggregate the entirety of the wealth of expertise and skill within *Transforming Musicology*. Captured metadata for all internal relationships and for each of the workflow stages results in a graph much richer than that produced through instance-level alignments alone.

[11] http://virtuoso.openlinksw.com/
[12] http://iiif.io/
[13] https://github.com/loris-imageserver/loris
[14] https://viaf.org/
[15] https://musicbrainz.org/

Although developed in the context of musicological investigation, the flexbility of the system - bar the niche ontologies themselves - has strong applicability across the Digital Humanities, breaking down barriers of information discovery between disciplines, supporting both innovative and traditional scholarship, and encouraging the re-use of tooling, data, and research methodologies.

Acknowledgments. This work was part of the UK AHRC *Transforming Musicology* project (AH/L006820/1). The authors acknowledge their colleagues on this project, especially Carolin Rindfleisch, and Richard Lewis.

References

1. Bechhofer, S., *et al*: Why Linked Data is not enough. Generation Computer Systems, 29, 2, pp. 599–611. Elsevier (2013)
2. Bechhofer, S., *et al.*: Computational analysis of the Live Music Archive. ISMIR (2014)
3. Bekiari, C., et al.:FRBR object-orientated definition and mapping from FRBRER, FRAD and FRSAD (version 2). International Working Group on FRBR and CIDOC CRM Harmonisation (2013)
4. Belhajjame,K., *et al*: Using a suite of ontologies for preserving workflow-centric research objects. Web Semantics: Science, Services and Agents on the World Wide Web (2015)
5. Cantum pulcriorem invenire: Conductus Database: `http://catalogue.conductus.ac.uk/#m-columnbrowser@||m-informationcontrol@url=html/home.php` (2013)
6. Crawford, T.: Early Music Online and The Electronic Corpus of Lute Music. MEI (2015)
7. Dix, A., *et al.*: Authority and judgement in the digital archive. DLfM (2014)
8. Fazekas, G.*et al*: An overview of Semantic Web activities in the OMRAS2 project'. Journal of New Music Research, 39(4):295311 (2010)
9. González-Beltrán A, *et al.* From peer-reviewed to peer-reproduced in scholarly publishing: The complementary roles of data models and workflows in bioinformatics. PLoS ONE 10(7): e0127612 (2015)
10. Jančovič, P., *et al.*: Automatic transcription of ornamented Irish traditional flute music using hidden Markov models. ISMIR (2015)
11. Khan, N., *et al*: BABY ElEPHãT - Building an analytical bibliography for a prosopography in early English imprint data. iConference (2016)
12. Missier, P., *et al*: Janus: from workflows to semantic provenance and Linked Open Data. IPAW (2010)
13. Mons, R.: Which gene did you mean?. BMC Bioinformatics, vol 6. p. 142 (2005)
14. Müllensiefen, D., *et al*: Recognition of leitmotives in Richard Wagner's music: chroma distance and listener expertise. ECDA (2014)
15. Oldman, D.: Contextual search design video: `https://www.youtube.com/watch?v=VUGMlDc9B5w` (2015)
16. Raimond, Y. *et al*: The timeline ontology. OWL-DL ontology (2006)
17. Raimond, Y. *et al*: The event ontology: Technical report (2007)
18. Rindfleisch, C: The Eternal Question to Fate, Surging up from the Fepth: Richard Wagner's Descriptions of his Leitmotives in Changing Contexts of Communication. RMA (2016)
19. World Wide Web Consortium: PROV-O: The PROV Ontology (2013)