An e-Research approach to Web-scale music analysis

BY DAVID DE ROURE¹,*, KEVIN R. PAGE¹, BENJAMIN FIELDS², TIM CRAWFORD², J. STEPHEN DOWNIE³ AND ICHIRO FUJINAGA⁴

¹Oxford e-Research Centre, University of Oxford, Oxford, UK
²Department of Computing, Goldsmiths University of London, London, UK
³Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL, USA
⁴Schulich School of Music, McGill University, Quebec, Canada

The growing quantity of digital recorded music available in large-scale resources such as the Internet archive provides an important new resource for musical analysis. An e-Research approach has been adopted in order to create a very substantive web-accessible corpus of musical analyses in a common framework for use by music scholars, students and beyond, and to establish a methodology and tooling that will enable others to add to the resource in the future. The enabling infrastructure brings together scientific workflow and Semantic Web technologies with a set of algorithms and tools for extracting features from recorded music. It has been used to deliver a prototype system, described here, that demonstrates the utility of LINKED DATA for enhancing the curation of collections of music signal data for analysis and publishing results that can be simply and readily correlated to these and other sources. This paper describes the motivation, infrastructure design and the proof-of-concept case study and reflects on emerging e-Research practice as researchers embrace the scale of the Web.

Keywords: music information retrieval; musicology; LINKED DATA

1. Introduction

Research disciplines from the sciences to arts and humanities are experiencing a change in practice in order to benefit from the wealth of data now available in digital form. This shift to an increasingly data-intensive research method, described in science as the fourth paradigm [1], is enabled by the new computational tools and techniques that characterize e-Science and e-Research. These enable researchers to examine data in new ways and to obtain new insights, effectively providing a new form of scientific instrument that can be thought of as a ‘datascope’: the socio-technical apparatus that takes us from ‘signal’ (the raw data from sensors and detectors of every form, capturing the physical world

*Author for correspondence (david.deroure@oerc.ox.ac.uk).

One contribution of 12 to a Theme Issue ‘e-Science: novel research, new science and enduring impact’.
in data) to new knowledge and understanding. Datascopes transcend scientific
disciplines and are equally useful in digital humanities, for example, in the study
of ancient documents [2].

In this paper, the signal is the vast number of digital recordings of music,
and the researchers who gain understanding are all those who study music in all
relevant disciplines. The datascope presented here is interesting in its own right
because no research instrument of this kind has been assembled before. It is also
interesting in the context of e-Research because it demonstrates the application
of a principled e-Research approach, both in terms of the underlying philosophy
and the architectural style.

1. A traditional approach has been to ‘warehouse’ data for analysis, thereby
locking it into specific projects and initiatives. A more sustainable and
scalable approach is to publish data (and even processes) as autonomously
created and independently evolving information sources that can be
composed without attempting to place them all in a predefined and rigid
framework. This promotes both sustainability and unanticipated reuse;
i.e. the outcomes of the research are not confined to the original scope and
duration of one research project.

2. While e-Science solutions have often adopted a service-oriented archi-
tecture, the systems described here are firmly based on the Web, i.e.
specifically in a resource-oriented architectural style. The Web itself
demonstrates the effectiveness of this approach, both in terms of scalability
and usability, and here we exercise it in an e-Research context.

These two principles have been rehearsed in earlier e-Science projects, for
example, in the e-Chemistry work of Taylor et al. [3]. They are exemplified now
by the use of LINKED DATA [4], an initiative whereby data are published on the
Web according to a set of conventions to maximize reusability: the movement
encourages a Semantic Web built upon Hypertext Transfer Protocol (HTTP)
Uniform Resource Identifier (URI) that are published, linked and retrieved using
resource description framework (RDF) and the SPARQL Protocol and RDF
Query Language (SPARQL) query language. While the LINKED DATA movement
is growing, it has not yet gained significant traction in scientific data [5]. Hence,
this paper also provides an investigation of LINKED DATA in an e-Research context.

The next section describes our initiative to analyse large amounts of music
information, and a case study is then discussed in §3. The architectural design
and implementation are presented in §4, and this is followed by a discussion and
conclusion in §5.

2. Analysis of large amounts of music information

(a) Background

There is now a tremendous volume of digital audio recordings available commer-
cially in private collections and online, covering many types of music. While most
prior analytical research work has focused primarily on Western popular and
‘classical’ music, this new dataset includes a wide variety of music from all over
the world, from many time periods and includes folk, classical, contemporary,
improvised and live music. For example, the Internet archive collection contains approximately 18,000 h of audio, including a substantial collection of live concert recordings (some 66,000 pieces), and represents a rich source for analysis that has hitherto been impossible. Analysis of this resource offers many benefits to music scholars, ranging from classical work recognition and genre classification to identification of national styles and more comprehensive study of ethnic music. In combination with other resources, it could enable research questions to be answered relating to the evolution of music over time and over geography.

Various algorithms and tools for extracting features from recorded music are available to support this analysis. These have been developed by the music information retrieval (MIR) and computational musicology (CM) communities over the last decade and evaluated through a series of annual international events called the Music Information Retrieval Evaluation eXchange (MIREX) [6]. The ability to analyse music directly in audio format is an important development: for example, in the past, most music structural analyses have been conducted using only those musical scores that were readily available, especially for European classical music, and the new audio-based information offers novel perspectives to music research, especially for ethnomusicologists, where no scores exist for many music cultures. The technical expertise needed to analyse music in audio format has prevented most music researchers from dealing with the actual performance of the music. With the recent revolution in MIR and CM research, many new tools and algorithms have been developed to analyse and to visualize music audio.

To tackle this scale of analysis, our *Structural Analysis of Large Amounts of Music Information* project (http://salami.music.mcgill.ca/) has adopted a new approach and is illustrated in figure 1. The algorithms chosen, modified and/or developed are being trained and evaluated using a set of ground-truth data that are based upon over 1000 exemplars created by trained musicologists.
The computational infrastructure for this scale of analysis makes use of a dataflow engine together with supercomputing time at the National Centre for Supercomputing Applications. The dataflow engine, MEANDRE [7], is an open-source dataflow execution framework designed to simplify the running of large-scale data mining/analysis applications on high-performance computing clusters and it stores the operational data of each session run in RDF making it easier to acquire and integrate the provenance data. An ontology for music structure is also under development to facilitate use of analyses, and the analysed data are being published using a LINKED OPEN DATA approach. This is based on the foundational work of Raimond et al. [8].

This initiative goes beyond current tooling and approaches in MIR. Some MIR systems have begun to incorporate data management and interoperability techniques: the Networked Environment for Music Analysis (NEMA) system [9] (used to operate MIREX 2010) adopts a service-oriented approach of subsuming existing MIR tools as services, but is limited to those that can be aligned with its Java data structures; the JMIIR suite uses the ACE XML document type definitions [10], adoption of which is therefore a prerequisite for interoperability; GNAT and GNARQL software tools [11] use the Music Ontology [8] (a key ontology used in this paper) to annotate only personal collections of music; while Henry [11], the SONIC VISUALISER and ANNOTATOR tools [12] and their VAMP audio analysis software plugins also use the Music Ontology, Feature Ontology and associated ontologies for import and export of data using the RDF model. However, these systems could be characterized as traditional MIR solutions that employ Semantic Web technologies rather than a resource-oriented architecture to support MIR research.

(b) Music information retrieval methodology

The following three steps broadly describe the process an MIR researcher may follow, the issues they raise and how each might be assisted by LINKED DATA and Semantic Web technologies:

1. Assemble a collection of audio input. To evaluate an algorithm, the researcher must acquire a wide selection of signal—typically digital audio files—for the algorithm to process. Music recordings are often restricted from free exchange among researchers, either explicitly through copyrights or implicitly through the high overheads of managing detailed and intricate licensing. Even when audio data are freely available and distributable, a difficult balance must be found to avoid ‘over-fitting’ of algorithms to a particular set of signals: whilst a widely shared, understood and reusable collection is critical for comparative evaluation, tuning an algorithm to such a collection during development (knowing it will be the benchmark) is likely to affect performance detrimentally against more randomly selected input (i.e. real-world tests). It is therefore useful to create and modify large collections of audio data quickly and flexibly and to share them between researchers for comparative evaluation. Linking existing metadata for audio files and basing collection generation on this information are desirable for quickly trialling an algorithm against particular musical facets (e.g. a particular period and style derived from the information about the composers).
2. **Apply the algorithm to the audio input.** There are many MIR systems that enable an algorithm to be applied to the signal. More recently, some systems have begun to adopt practices and tools from the scientific workflow community, for example, the MEANDRE workflow enactment system. Any such system must be able to accept an input collection and apply the algorithm across it. Where institutionally restricted collections of signals are in use, a system must match local audio files to any abstract, metadata-based, collection descriptions.

3. **Publish and evaluate algorithm output.** The MIR community has a 7 year history of comparative evaluation in the MIREX exercise; the most recent (2010) MIREX adopted a derived framework derived from MEANDRE in order to execute the algorithms under test [9]. More generally, evaluation of results requires a common framework into which analytical output can be aligned and published for comparison, rather than making with data structures inherited from the development tool or the environment a researcher was using. As faster computational resources become more readily available and can be applied to MIR tasks, the opportunity to undertake analysis on an ever-greater scale brings with it the associated problems of managing ever-greater quantities of result data. Links from results back to recorded signal (and audio file artefacts) and capturing provenance are equally important: a single algorithm is not normally sufficient to make a definitive assertion, e.g. to classify a recording as jazz. For this reason, it is important that the resultant data can be used as input for creating derivative collections of input for further MIR analysis such that information extracted from multiple algorithms can be combined and refined.

It is interesting to note how metadata exchange and integration helps the MIR researcher operate in a domain in which there are restrictions on copying the signal data, but where other versions of some of that data may be accessible. Restrictions on the distribution of actual audio files can be accommodated through separate description of collections as metadata and correctly modelling the relationship between a work and derivative artefacts (e.g. distinguishing between a work, a performance of the work, recordings of the performance, published media of the recording). Metadata exchange can then occur independently and be cross-referenced against any institutional or other private archive of audio; that is, inputs to algorithms can be described using metadata when there is an expectation that equivalent collections of signal are available on each of the systems where the algorithm is to be executed.

3. **The genre recognition demonstrator**

While the principles and design described here can be applied to all MIR systems, a specific use case has been developed for demonstration purposes. In this scenario, the researcher is investigating a correlation between the country of residence of a recording artist and the genre of a performance, that is, the labelling according to ‘style’, e.g. country, jazz, rock, baroque, as classified
by an MIR algorithm. Signal collections are derived from the country of performers, descriptive metadata is gathered and published, then genre analysis and integration of collection and result metadata enables the user to ask ‘how country is my country?’.

This prototype, known as ‘Country/Country’, embraces the principles of the LINKED DATA movement, which encourages a Semantic Web built upon HTTP URIs that are published, linked and retrieved using RDF and SPARQL. By employing new RDF encodings for collections and results that use existing ontologies (including the Music Ontology, GeoNames (http://www.geonames.org/ontology/), Provenance Vocabulary [13] and Object Reuse and Exchange (ORE) [14]) and by deploying a LINKED DATA audio file repository and services for publishing collections and results, we present a proof-of-concept system that addresses the problems outlined earlier.

The components of the system align with the steps in §2 with the addition of a pre-step. The generic purpose of each service, and the specific implementation in Country/Country, are as follows (further implementation detail is provided in §4):

0. An audio file repository that serves audio files and LINKED DATA about the audio files using HTTP. Using the Music Ontology, the relationship to the track it is a recording of, and the ‘definitive’ URI for that track, is asserted in the LINKED DATA. For the public demonstrator, a subset copy of the Jamendo free music collection (http://www.jamendo.com/) has been used, and the URIs are minted by the Jamendo LINKED DATA service at dbtune (http://dbtune.org/jamendo/).

1. A collection builder web application enables a user to publish sets of tracks described using RDF. The backend uses SPARQL to build collections and takes advantage of links between datasets. An optional second stage of the collection builder takes a collection and ‘grounds’ the constituent tracks against available recordings of those tracks by posing SPARQL queries to audio file repositories. The Jamendo service incorporates links to geographic locations as defined by GeoNames (http://www.geonames.org/ontology/), so the collection builder can identify all of the tracks offered by Jamendo recorded by artists from a specific country. In the case of Country/Country, we ground a country-derived collection against our audio file repository of locally available signal.

2. The analysis is performed by a NEMA genre classification workflow. The myExperiment [15] scientific collaborative environment has been extended to support the MEANDRE [7] workflow used by NEMA. myExperiment has also been modified to accept the collections RDF published in step 1 and marshal the target tracks contained within the analysis workflow. Within the MEANDRE-based genre classification workflow, a head-end component has been written to dereference each track URI passed to the workflow and, using the LINKED DATA published by the signal repository, retrieve both the local copy of the audio file and the reference to the original Jamendo identifier. This URI persists through the genre analysis workflow until it reaches a new tail-end component where the analysis is published using RDF; including links back to the Jamendo URI.
3. **A RESULTS VIEWER** Web application retrieves the collections RDF from step 1 and results RDF from step 2, by cross-referencing them via the URIs used throughout the system. The user can look for trends in genre classification within and between collections. Results can be combined for comparison purposes using existing and new collections, and observed patterns can then inform the creation of new collections. To illustrate that further links can easily be made to existing datasets and inform derivative collection generation, relevant associations from other LINKED DATASETS are shown; e.g. artists of the same genre and country from DBpedia [16] and the BBC for a particular analysed track.

4. **System design**

The services form a highly decentralized, distributed, loosely federated and scalable resource oriented architecture [17] shown in figure 2: interactions between services occur over HTTP and involve the exchange of representations of resources identified globally by URIs. While the sequence above is repeated throughout the paper to explain the utility of the services in the context of the use case, there is no requirement for services to interact in this, or any other, specific order. Since this is a proof of concept, each service is neither the singular nor definitive implementation of its type. For speed of development and clarity of explanation, the instances of each service presented here are limited examples, but in a true Web of data, there would be many providers of all service classes.

(a) **Audio file metadata and repository**

The starting point for most music analysis tasks is the selection of input data to be processed by the algorithm under development. In our prototype system, we focus on the provision of audio signal data in the form of MP3 files, but the technique could equally be applied to symbolic source material such as the Musical Instrument Digital Interface (MIDI).

There are many bases upon which a researcher might assemble and manage a selection of input signal data; often, this may be down to the practicalities of local availability of physical media, or freely accessible remote collections. Such limitations of collections cause not only validity issues such as over-fitting algorithms to test data, but also preclude the discovery of novel research techniques and results that might be expected when analysing the massive and increasing digital corpus. In this demonstrator, we show how metadata can be used to automate the assembly and management of larger, distributed and more dynamic collections. While limited metadata is often available through mechanisms such as the ID3 standard used in conjunction with the MP3 audio file format, this is usually no more than a simple string tag and is limited in scope to the specific audio file in question—here we apply Semantic Web techniques to retrieve and combine metadata from various sources, both directly and indirectly related to the signal data. In addition to the attributes of a particular audio file (sampling rate, play length, etc.), this metadata can capture the lineage between the file and more abstract notions relating to artists and their work, including concepts such as distribution (copying of digital artefacts), encoding (digitization or compression), recording (of a performance of a work) and composition (of the work by an artist).
A powerful and flexible feature of the Semantic Web is the ability to distribute metadata across the Web while maintaining a common foundation in the underlying (RDF) model and, as is often desirable, shared ontologies. For example, metadata about an artefact such as an audio file can be maintained on a different Web server than both the audio file itself and other distinct sources of metadata, but when required, the metadata can be dynamically combined to form a coherent statement of information about, and with, the audio file. Building and maintaining this Web of distributed information is a key motivator for the LINKED DATA community.

Phil. Trans. R. Soc. A (2011)
In this system, the first links to this Web are made by an audio file repository, which serves both MP3 audio files and LINKED DATA about the audio files. While this typically represents a generic collection of audio files to which there may be open or restricted access, our public demonstrator has been created by amassing a subset of the freely available Jamendo collection. The repository consists of an Apache Web server that has been configured to conform to Representational State Transfer (REST) [18] and LINKED DATA principles [19] such that:

— the primary resources are AudioFiles\(^1\) as described by the Music Ontology [8]. URIs are minted for these resources within the namespace of the repository, e.g. http://repository.nema.linkedmusic.org/audiofile/100002;
— if a client fetches the URI representing an AudioFile non-information resource and uses the HTTP Accept header to request audio/* (e.g. audio/mpeg), then the server issues a 303 redirect to the audio signal file (an information resource), which the client can then download;
— if, through the HTTP Accept header, a client requests application/rdf+xml, then the server issues a 303 (‘see other’) redirect to a LINKED DATA RDF file (another information resource) containing metadata pertaining to the AudioFile; and
— the RDF sub-graphs for each AudioFile are written to a 4store (http://4store.org/) triplestore (a purpose-built RDF database) that provides a SPARQL query endpoint.

A second motivation for populating the audio repository with music from the Jamendo label is to use the LINKED DATA endpoint for Jamendo available at dbtune (http://dbtune.org/jamendo/). This in turn enables LINKED DATA publication (figure 2(0)), served when a client requests RDF (as above), and using the Music Ontology to assert that:

— an audio file resource found in the repository is an instance of an AudioFile;
— each AudioFile in the repository encodes a specific linked signal instance as defined in the Jamendo LINKED DATASET (where signal is the concept as defined by the Music Ontology); and
— a specific Track instance in the Jamendo RDF graph is encoded by each AudioFile in the repository.

While the open licensing of Jamendo enabled a public demonstrator to be built, there is no fundamental requirement for the audio files to be sourced from the same provider as the LINKED DATA—as shown in later sections, the aim is to encourage the opposite. For example, the audio files could be transcodings from a private collection with access restricted on an institutional basis, while album metadata would be linked from the Musicbrainz endpoint (http://dbtune.org/musicbrainz/).

\(^1\)Capitalized terms throughout this paper refer to concepts defined in ontologies, e.g. the Music Ontology.

Phil. Trans. R. Soc. A (2011)
Figure 3. Screenshots of the Country/Country component user interfaces: (a) the collection builder, (b) the genre analysis workflow in myExperiment and (c) the results viewer. (Online version in colour.)

(b) Collection builder Web application

(i) Creating collections

While provision of a LINKED DATA audio file repository was a necessary building block in construction, the Country/Country prototype, a key motivation for this approach is to free MIR researchers from datasets that are directly derived from specific signal repository contents. Dynamic collections spanning multiple repositories could instead be selected using criteria relevant to the research being undertaken, whether from within or outside the MIR domain; earlier experimental results could be fed back into this process as further criteria for creation of derivative collections.

For purposes of demonstration, a simplified use case is considered where a researcher wishes to investigate the possible correlation between the genre of a performance as detected by an MIR algorithm and the domicile of the performing artist. The collection builder Web application (figure 3a) provides the user with an interface to create collections from the entire Jamendo community, rather than being limited to the subset of signal served by the audio file repository. As the user selects filters, SPARQL queries are built up beneath the user
interface, using concepts within and beyond Jamendo and the Music Ontology to query for *signal* instances. An illustrative SPARQL query can be found in the appendix.

Once the user has applied sufficient filters to achieve their desired criteria and a SPARQL query constructed to enact it, the user may ‘publish’ their collection. This takes the form of RDF whereby the collection builder mints a URI for the RDF collection in the http://collections.nema.linkedmusic.org/ namespace and asserts the collection as an ORE *Aggregate* of signal, where the signals are URIs from the Jamendo namespace that match the SPARQL query. It then uses the Provenance Vocabulary to record the SPARQL query used and asserts user-specified additional metadata including authorship and description.

(ii) *Grounding collections with audio files*

The collections described in the previous section have been selected by criteria unbound by an audio file repository, but they only contain the abstract notion of signal as defined by the Music Ontology; to be used as input by an MIR algorithm, they must be grounded as AudioFiles that encode corresponding signal in one or several signal repositories.

A second stage of the collection builder Web application enables a user to do just this. By querying the SPARQL endpoint provided by the audio file repository (§4a) for the signal URIs aggregated in the abstract collection, a second RDF aggregation is published: a URI is minted for this grounded collection, an ORE aggregate is asserted this time containing AudioFile URIs from the audio file repository (§4a) that encode signals from the existing abstract collection, and the Provenance Vocabulary expresses the relationship between the grounded collection and its abstract precursor.

It should be noted that an AudioFile collection might not be a complete grounding of a signal collection: coverage is restricted to that of the audio file repository (or repositories) available. On the other hand, multiple corresponding AudioFiles may be available and encoded in the grounded collection, whereby the appropriate repository would be selected when the AudioFile required is determined by network speed, locality or access restricted by licence.

(c) *MEANDRE workflow and results repository*

The MEANDRE data-intensive flow framework has been adopted as the workflow enactment engine at the core of the NEMA that has been used as the submission and evaluation framework for MIREX 2010 (http://www.music-ir.org/mirex/wiki/2010:Main_Page). The heart of the NEMA system design is an extensible Java data model that incorporates MIR data structures from existing tools such as jMIR [20] and the SONIC tools [12]; in combination with the distributed execution environment of MEANDRE, this allows the NEMA system to host and run MIR workflows authored in a wide variety of existing languages and environments [9].

The MIR stage of the Country/Country prototype adopts an existing MEANDRE workflow that performs genre and mood analysis, i.e. it takes an audio signal as input, and through a workflow of feature extraction and a number of trained classifiers (e.g. Classification and Regression Tree (CART), J48 decision tree, Linear Discriminant) it provides a weighted ranking of genre (e.g. country,
baroque, jazz, rock) and mood (e.g. aggressive, wistful, cheerful) for each audio signal. If an analysis has been previously run on a given signal with the same parameters, then the results are already available: this is one way in which the method scales, as the most popular queries will be answered without any (audio) processing.

(i) MEANDRE components

Each component in a MEANDRE flow is encapsulated by a Java object, and to integrate with the LINKED DATA services provided by Country/Country, the ‘head’ and ‘tail’ components of the flow have been modified such that:

— The head component, which retrieves and passes an audio signal to the feature extractor, has been adapted to parse a LINKED DATA AudioFile URI—such as the one provided by our audio file repository (§4a)—as its input. The component dereferences this non-information resource twice: once with the audio/mpeg HTTP Accept header to get the audio signal file and again requesting application/rdf+xml to retrieve the LINKED DATA pertaining to the AudioFile.

— The RDF sub-graph retrieved from the audio file repository is stored using an in-memory JENA (http://jena.sourceforge.net/) model so that the URIs can persist through the flow. This maintains the crucial links between the audio signal retrieved from the repository and processed by the flow, and the global identifiers—the URIs—of the signal of which the AudioFile is an artefact, and—via signal related concepts such as Artist and Track—LINKED DATA sources such as Jamendo.

— The tail component outputs the weighted rankings by genre and mood from the classifiers: the results of the analysis. Because the RDF sub-graph includes concepts from the Music Ontology for both global identifiers (e.g. for signal) and local artefacts (AudioFile), we can distinguish between these when recording results. Genre, for example, is a concept applied to a signal, for which the AudioFile is a digital artefact of a signal (that in turn encodes a Performance).

— Analysis is performed on a sampled frame-by-frame basis within the workflow, so output is written both as a comma-separated values file containing detailed classifier values for each frame, and as a LINKED DATA RDF model with the average analysis for the whole Performance (i.e. per AudioFile).

— The RDF result graphs are also inserted into a 4store triplestore to provide a SPARQL query endpoint.

(ii) Results repository resource description framework

The tail component of the workflow uploads output from the analysis to a Results Repository (figure 2(3)). The fundamental resources in the repository are ORE Aggregations containing Associations (as defined in the Music Similarity Ontology [21]), where the Aggregation of Associations corresponds to the results from a single classifier analysing a single AudioFile. URIs are minted for these associations in the http://results.nema.linkedmusic.org/ namespace.
For example, output from the genre classifiers is modelled using a locally declared \textit{GenreAssociation} subclass of Association, which has as its subject a signal instance (derived from the AudioFile via the audio file repository \textit{Linked Data}), and as its object a \textit{MusicGenre} instance as defined by the DBpedia Ontology.

Further Provenance Vocabulary is used to record the \textit{MEANDRE} flow execution instance that performed the analysis (\textit{createdBy}), the classifier within the flow (\textit{usedGuideline}) and the AudioFile input to the analysis (\textit{usedData}, as distinct from the parent Performance of which the AudioFile is a derivative artefact). The Comma-separated Values (CSV) file containing frame-by-frame analysis is linked using the Opaque Features File Ontology (http://purl.org/ontology/off/), which describes references to feature data considered too densely numerical for efficient ontological representation.

\textit{(d) myExperiment workflow management}

The myExperiment \cite{15} Web-based virtual research environment provides discovery, sharing and management of workflows and associated Research Objects throughout their lifecycle, providing specific support for TAVERNA workflows.

Support for \textit{MEANDRE} workflows, as used by the NEMA system and the Country/Country prototype, has been added to myExperiment. This includes a preview page for \textit{MEANDRE} flows, the same ability to share and manage \textit{MEANDRE} flows as for TAVERNA and functionality to enact the flow on a specified \textit{MEANDRE} flow server (figure 3\textit{b}). The underlying implementation stores \textit{MEANDRE} archive units (complete self-contained \textit{MEANDRE} workflows including executable components and workflow metadata) within the myExperiment system.

The myExperiment Application Programming Interface (API) has also been extended to support importing collections from the Country/Country collection builder (\S4\textit{b}). This new API method takes the URI of a grounded collection as its argument; when accessed, myExperiment loads the Collection metadata and makes it available to a user as potential input to a workflow. Should the user then apply the collection to the Country/Country genre analysis workflow, myExperiment will iterate through the collection and enact the workflow for each AudioFile URI within it (each AudioFile URI is then dereferenced within the workflow; see \S4\textit{c}).

A link utilizing this API call is appended to the end of the collection builder grounding process so that a user can quickly and simply move from collection maintenance to application of the collection to workflows in myExperiment.

\textit{(e) RESULTS VIEWER Web application}

The final service provided as part of the Country/Country prototype system is a Web application that allows a researcher to view the analysis results, to cross-reference against collections and to combine the analysis with other \textit{Linked Data} sources. More than any other component, the \textit{RESULTS VIEWER} is a proof of concept that highlights only a select number of the many possible data sources and combinations.

The \textit{RESULTS VIEWER} demonstration implementation (figure 3\textit{c}) begins by combining two \textit{Linked Datasets}: it takes a collection (as created in the collection builder, \S4\textit{b}) and queries the Result Repository SPARQL endpoint (\S4\textit{c}) matching Association results for signal contained in the collection(s).
The demonstrator is focused on our country-centric genre analysis scenario: using country-derived collections cross-referenced with result data, a number of statistics and visualizations pertinent to this scenario are calculated and rendered including:

— for a collection (and comparison of multiple collections): the number of signals, the number of signals that have been grounded in an audio file repository, the number of classifiers that were run on any of the signals according to a Results Repository and the numbers of results (genre associations) available for each workflow enactment of the classifier;
— for each classifier over a collection: the songs (by artist and title) that are most and least weighted for each genre, a pie chart taking the highest genre weighting for each signal and a pie chart showing average weightings for each genre over the collection; and
— for each signal in the collection: a full listing of genre weightings from each classifier, a playback page that retrieves and plays the AudioFile (using LINKED DATA from the audio file repository) and the frame analysis data (using the data links from the Results Repository RDF) and references to other relevant information from LINKED DATA sources, demonstrating the further potential of LINKED DATA in bringing together a wide variety of information sources.

The first example takes the highest weighted genre for a given artist or collection and links to other artists in DBpedia who perform in the same genre and are also from the same country. This illustrates how it is possible to link between imperfectly aligned datasets: not only do GeoNames (and the linked Jamendo dataset and the Country/Country collections) and DBpedia use different ontologies for countries, but artists in DBpedia can also be associated with a wide variety of geographic coverages (town, region, country, etc.) and through various relationships (residence, place of birth, etc.). To overcome this, geographic entities below the level of country in DBpedia have been asserted ‘sameAs’ specific features in GeoNames—in other words, even though it is conceptually incorrect to align the ontologies at the level of country, it is possible at other levels (e.g. cities and towns).

A SPARQL query to DBpedia for a list of geographical locations associated with all artists of a specific genre can then be cross-referenced against their sameAs features in GeoNames, which can finally be culled by the GeoNames country they are located in (as used by the Country/Country collections). Although the relationship between artist and country in DBpedia can be one of several types, the common RDF model allows us to process them all. The second example takes this list of artists and, using the provided sameAs assertions, links to the same artists on the BBC Music website (http://www.bbc.co.uk/music/).

5. Conclusions and future work

Researchers in the field of MIR are confronted with problems beyond the design and implementation of systems and algorithms for retrieving information from music. The music recordings over which analysis would be expected to occur
are often restricted from exchange among researchers, either explicitly through copyrights or implicitly through the high overheads of managing detailed and intricate licensing. As increasingly vast quantities of audio data are digitized, their entanglement with rights management will only make the curation and distribution of ever larger datasets a more complicated and time-consuming task. Even when audio data are freely available, a difficult balance must be found between the need for comparative evaluation of approaches using widely shared, understood and reusable datasets, and the avoidance of over-fitting an algorithm during development when a specific dataset is repeatedly used for testing.

Evaluation also requires a common structure into which analytical output can be placed for comparison, rather than the data structures inherited from the development tool or environment a researcher happened to be using. As faster computational resources become more readily available and can be applied to MIR tasks, the opportunities to undertake analysis on an ever greater scale [22] bring the associated problem of managing ever greater quantities of result data. In this context, the Country/Country prototype demonstrates the utility of Semantic Web technologies: the consistent use of globally unique identifiers (in the form of URIs) that can persist within and between systems; a resource-oriented architecture that enables highly distributed, lightweight and dynamic services when publishing data; and a common underlying model in RDF and shared ontologies for information exchange and the power of merging distributed information through a Web of LINKED DATA.

Future work will more completely and accurately model the data and processes within analysis workflows—‘black boxes’ within the current system, but which are the focus of any MIR researcher’s work and interest. While MEANDRE is nominally underpinned by an RDF data model, the structure of this model is as yet insufficient for direct publication as LINKED DATA (extensive use of string literal key/value pairs limits the opportunities for linking). Furthermore, procedures such as collection building are themselves workflows, and as the quantity of LINKED DATA available for collection building increases, the value of applying workflow techniques and sharing environments can only grow. While this research has demonstrated how RESTFUL and LINKED DATA techniques enable the distributed serving and separation of content and metadata, a future implementation should demonstrate how standard HTTP access and authentication mechanisms can take advantage of this separation for the purpose of adhering to digital rights restrictions on audio content.

Our case study has allowed us to obtain feedback from the MIR, LINKED DATA and e-Research communities [23] and to illustrate how even a relatively limited linking of Semantic Web data sources can provide an MIR researcher with a far greater flexibility when selecting input sources than previously available. When links to the RDF graph are maintained through an analysis workflow, the results published as LINKED DATA can be quickly, easily and usefully cross-referenced with other results, signal collections and further sources of data beyond the obvious day-to-day purview of the researcher.

While the demonstrator embodies a specific analysis (genre) and collection selection (by nation) in a basic use case, the system illustrated here is more generic and widely applicable to the MIR community and beyond. Our case study has proven to be a very useful vehicle to exercise and hone our design principles: we believe that this approach—of a distributed Web of independently
Web-scale music analysis

3315

evolving information sources linked by a shared representational and addressing framework—has potential for general application to e-Research. The common data model of RDF extends a myriad of possibilities for linking models and categorizations within and between research disciplines; designing the infrastructure to deliver this data to the researcher is an opportunity to provide interfaces that embrace these domain requirements [24].

One should not overstate the current availability of LINKED DATA, for while there are plentiful opportunities for improving the lot of researchers using the current sparse link density, information exposed as LINKED DATA is but a tiny fraction of that available on the World Wide Web—the document Web. Herein lies something of a bootstrapping problem that can be tackled by the easy, inconspicuous and simple data publishing techniques illustrated here.

The tools presented offer increased automation and simplification of day-to-day tasks, greater impact of results through easy access and in turn more frequent reuse and validation by peers. While a researcher may not immediately or directly recognize the benefits of idealized LINKED DATA principles, such practical benefits must surely be an attractive motivation that could kick-start a virtuous circle of reuse and automation: one researcher’s results can form the basis for another’s input collection, so data and techniques can be combined, the Web of LINKED DATA grows and the scale of reuse and automation grows further.

This work was carried out through the Structural Analysis of Large Amounts of Musical Information project, part of the international Digging into Data challenge and was funded by the JISC Digitisation and e-Content programme, the National Science Foundation (grant nos IIS 10-42727 and IIS 09-39253 and by SSHRC (869-2009-0001)); it builds on previous work funded under the Networked Environment for Musical Analysis project funded by the Andrew W. Mellon Foundation. The Web interface for the Country demonstrator and the myExperiment extension for MEANDRE were developed by Bart J. Nagel and Gianni O’Neill, respectively, at University of Southampton, UK.

Appendix A

The query below is used to return details of tracks recorded by artists from the country of Belgium, where the location of an artist is asserted in the Jamendo data, but the country of that artist’s location is encoded by GeoNames.

PREFIX geo: <http://www.geonames.org/ontology#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX mo: <http://purl.org/ontology/mo/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT * WHERE {
  ?artist a mo:MusicArtist ;
  foaf:name ?artistname ;
  foaf:based_near ?basednear .
  OPTIONAL { ?basednear geo:inCountry ?country . } 
  ?record

Phil. Trans. R. Soc. A (2011)
a mo:Record ;
foaf:maker ?artist ;
mo:track ?track ;
dc:date ?recorddate ;

?track
dc:title ?trackname ;
mo:track_number ?tracknumber .

?signal
mo:published_as ?track .

References


