An e-Research Approach to Web-Scale Music Analysis

David De Roure[1], Kevin R. Page[1], Benjamin Fields[2], Tim Crawford[2],
J. Stephen Downie[3], Ichiro Fujinaga[4]

[1] Oxford e-Research Centre, University of Oxford, UK

[2] Department of Computing, Goldsmiths University of London, UK

[3]Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, US

[4]Schulich School of Music, McGill University, Canada

Abstract

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 which 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 prototypical 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.

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[11], is enabled by the new computational tools and techniques that characterise e-Science and e-Research. These enable researchers to examine data in new ways and 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 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 [22].

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 like this 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 system architecture and the underlying philosophy:

1. While e-Science has traditionally adopted a service oriented architecture, the systems described here are firmly based on and in the Web, i.e. they are resource-based. The Web itself demonstrates the effectiveness of this architectural style both in terms of scalability and usability, and here we exercise this approach in an e-Research context.

2. While 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. This promotes unanticipated reuse; i.e. the outcomes of the research are not confined to the original scope and duration of one research project.

These two principles have been rehearsed in earlier e-Science projects, for example in the e-Chemistry work of Frey [23]. They are exemplified now by the use of Linked Data, an initiative whereby data is published on the Web according to a set of conventions to maximise reusability: the movement encourages a Semantic Web built upon HTTP URIs that are published, linked, and retrieved using Resource Description Framework (RDF) and the SPARQL query language [5]. While the Linked Data movement is growing, it has not yet gained significant traction in scientific data [3]. Hence this paper also provides an investigation of Linked Data in an e-Research context.

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

2 Analysis of Large Amounts of Music Information

2.1 Background

There is now a tremendous volume of digital audio recordings available commercially in private collections and online, covering many types of music. While most prior analytic 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 ~18,000 hours of audio including a substantial collection of live concert recordings (~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) [1]. 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 to analyse and to visualise music audio have been developed.

To tackle this scale of analysis, our *Structural Analysis of Large Amounts of Music Information* project [1] has adopted a new approach, illustrated in Figure 1. The algorithms chosen, modified and/or developed are being trained and evaluated using a set of ground-truth data based upon over one thousand 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 Center for Supercomputing Applications (NCSA). The dataflow engine, Meandre, 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. A standardised ontology for music structure is also under development, to facilitate use of analyses, and the analysis data is being published using a Linked Open Data approach. This is based on the foundational work of Raimond[18].

This initiative goes beyond current tooling and approaches in Music Information Retrieval. Some MIR systems have begun to incorporate data management and interoperability techniques: the Networked Environment for Music Analysis (NEMA) system[24] (used to operate MIREX 2010) adopts a Service Oriented approach of subsuming existing MIR tools as services, but is limited to those which can be aligned with its Java data structures; the jMIR suite uses the ACE XML DTD[15], adoption of which is therefore a

[1] http://salami.music.mcgill.ca/

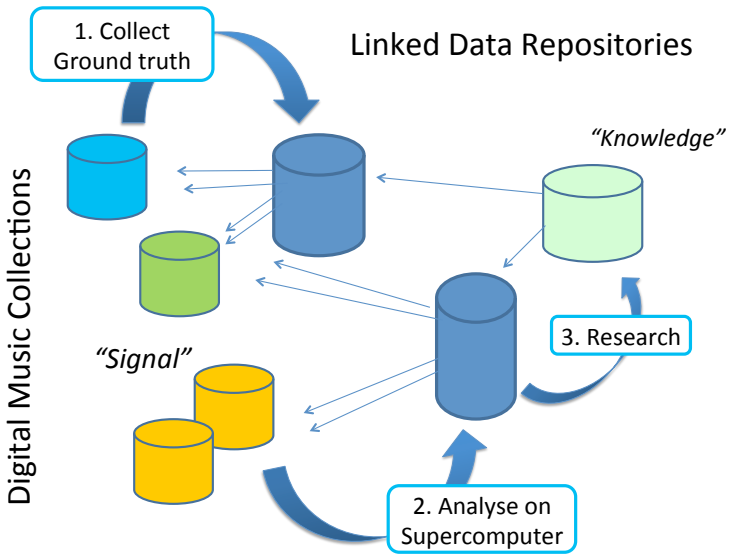

Figure 1: Digital music collections provide the signal which is analysed by experts (1) and, based on this, by machine (2) in order to publish new Linked Data resources that are used to support musicologists (3). New signal and new results can be added by the community.

prerequisite for interoperability; GNAT and GNARQL[19] use the Music Ontology (a key ontology utilised in this paper) to annotate only personal collections of music; while Henry[19], the Sonic Visualiser and Annotator tools[6] and their VAMP plugins also use the Music Ontology, Feature Ontology, and associated ontologies for import and export of data using the Resource Description Framework (RDF) model; however, these systems could be characterised as traditional MIR solutions that employ Semantic Web technologies rather than a Resource Oriented Architecture to support MIR research.

2.2 MIR 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 amongst researchers, either explicitly through copyrights or implicitly through the high overheads of managing detailed and intricate licensing. Even when audio data is 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 re-usable 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. Restrictions on the distribution of actual audio files can be accommodated through the separate description of collections and correctly modelling the relationship between artefacts (e.g. distinguishing between a work, a performance of the work, recordings of the performance, and published media of the recording); metadata exchange can then occur independently and be cross-referenced against any institutional or other private archive of audio. Linking existing metadata for audio files and basing collection generation on this information is 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 signal. More recently some systems have begun to adopt practices and tools from

the scientific workflow community, for example the Meandre workflow enactment system [14]. Any such system must be able to recognise an input collection and apply the algorithm across it. Where institutionally restricted collections of signal 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 seven-year history of comparative evaluation in the MIREX competition; the most recent (2010) MIREX adopted a Meandre derived framework for executing the algorithms under test [24]. More generally, evaluation of results requires a common structure into which analytic output can be published for comparison, rather than making do 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 representation of results 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.

3 The genre recognition demonstrator

The Linked Data movement encourages a Semantic Web built upon HTTP URIs that are published, linked, and retrieved using RDF and SPARQL. Employing new RDF encodings for collections and results that utilise existing ontologies (including the Music Ontology, GeoNames, Provenance Vocabulary, and Object Reuse and Exchange), 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 in the previous section.

While the principles and design described here can be applied to all MIR systems, a specific use case known as "Country/Country" has been developed for demonstration purposes. Metadata describing signal collections derived from the country of an artist are gathered and published, then genre analysis and integration of collection and result metadata enables the user to ask: "how country is my country?" The components of the system align with the steps in the previous section with the addition of a pre-step. The generic purpose of each service, and the specific implementation in Country/Country, is as follows:

0. An Audio File Repository which serves audio files and linked data about the audio files using HTTP. Using the Music Ontology[18], 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 free-licensed Jamendo collection[2] has been used, and the URIs are minted by the Jamendo linked data service at dbtune[3].

1. A Collection Builder web application that 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[4], so the Collection Builder can identify all 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[24] genre classification workflow. The myExperiment[9] scientific collaborative environment has been extended to support the Meandre[14] workflows used by NEMA. myExperiment has also been modified to accept the collections RDF published in step 1) and marshal

[2] http://www.jamendo.com/
[3] http://dbtune.org/jamendo/
[4] http://www.geonames.org/ontology/

the target tracks contained within to 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 1) and results RDF from 2), cross-referencing them via the URIs used throughout the system. The user can identify trends in genre classification within and between collections. Results can be pooled and compared using existing and new collections and inform the creation of new sets. To demonstrate how further links can easily be made to existing datasets and inform derivative collection generation, relevant associations from other linked data sets are shown (e.g. artists of the same genre and country from DBpedia and the BBC for a particular analysed track).

4 System Design

The services form a highly decentralised and distributed, loosely-federated, and scalable Resource Oriented Architecture[20] 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 through 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 – in a true web of data there would be many providers of all service classes.

4.1 Audio File Metadata and Repository

The starting point for most music analysis tasks is the selection of input data for the algorithm under development to process. 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 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[7]. 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 ID3, 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.

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.

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[5]. The repository consists of an Apache web server that has been configured to conform to REST[8] and Linked Data principles[21] such that:

[5] http://www.jamendo.com/

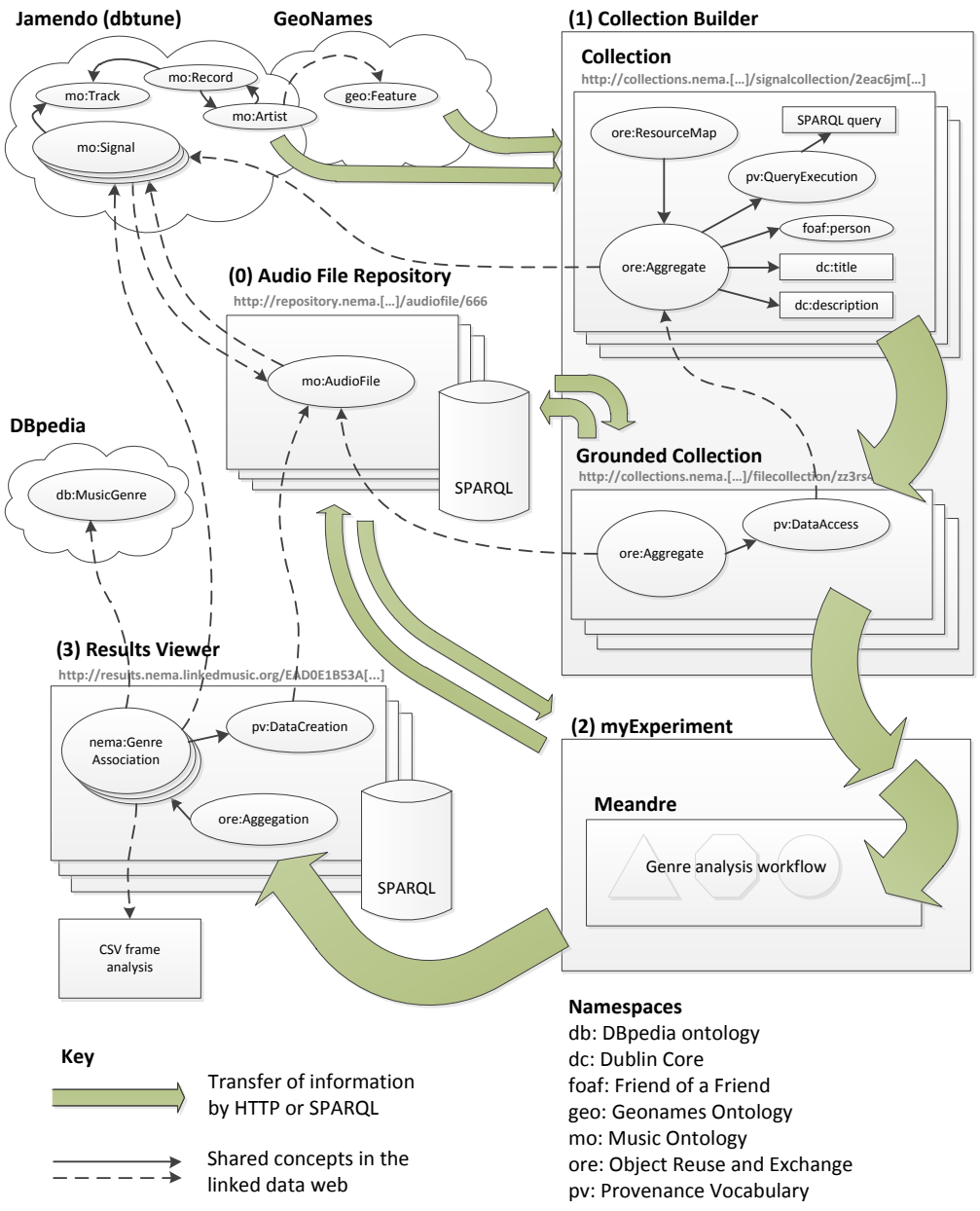

Figure 2: Country/Country system architecture with some connections in the linked data web overlaid

- the primary (non-information) resources are *AudioFiles*[6] as described by the Music Ontology[18]. 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 redirect to a linked data RDF file (another information resource) containing metadata pertaining to the AudioFile.
- the RDF sub-graphs for each AudioFile are written to a 4store[7] triplestore which provides a SPARQL query endpoint.

A second motivation for populating the audio repository with music from the Jamendo label is to utilise the Linked Data endpoint for Jamendo available at dbtune[8]. 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 data set (where Signal is the concept as defined by the Music Ontology).
- 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[9].

4.2 Collection Builder Web Application

4.2.1 Creating collections

While provision of a linked data Audio File Repository was a necessary building block in construction of the Country/Country prototype, a key motivation for this approach is to free MIR researchers from data sets 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 (e.g. country, jazz) as detected by an MIR algorithm, and the domicile of the performing artist. The *Collection Builder* web application (figure 3(i)) 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 (figure 2(2)), 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[13] *Aggregate* of Signal, where the Signals are URIs from the Jamendo namespace that match the SPARQL query. It then uses the Provenance Vocabulary[10] to

[6]Capitalised terms throughout this paper refer to concepts defined in ontologies, e.g. the Music Ontology.
[7]http://4store.org/
[8]http://dbtune.org/jamendo/
[9]http://dbtune.org/musicbrainz/

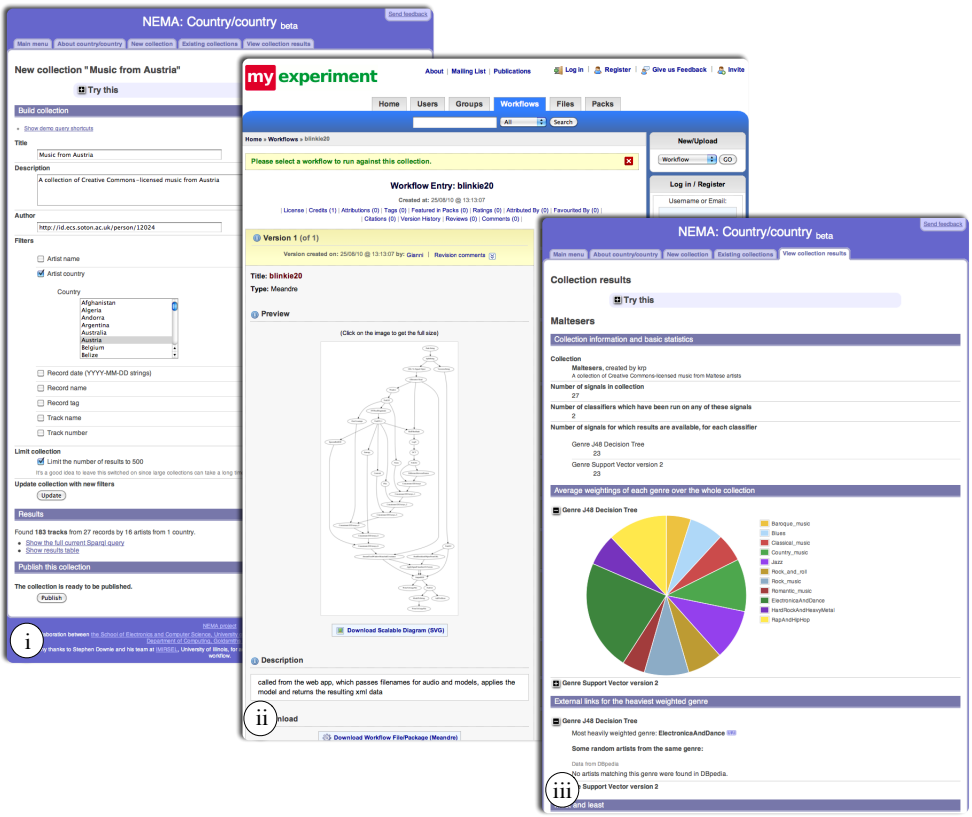

Figure 3: Screenshots of the Country/Country component user interfaces: (i) the Collection Builder, (ii) the genre analysis workflow in myExperiment, (iii) the Results Viewer.

record the SPARQL query used and asserts user specified additional metadata including authorship and description.

4.2.2 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 (section 4.1) 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 (section 4.1) that encode Signal 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 may 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 license restricted access.

4.3 Meandre Workflow and Results Repository

The Meandre data-intensive flow framework[14] has been adopted as the workflow enactment engine at the core of the Network Environment for Music Analysis (NEMA) system which has been used as the submission and evaluation framework for MIREX 2010[10]. The heart of the NEMA system design is an extensible Java data model that incorporates MIR data structures from existing tools such as jMIR[16] and the Sonic tools[6]; 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[24].

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. CART decision tree, J48 decision tree, Linear Discriminant) 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 then the results are already available from there: this is one way in which the method scales, as the most popular queries will be answered without any (audio) processing.

4.3.1 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 one provided by our Audio File Repository (section 4.1) – 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[11] 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 frame-by-frame basis within the workflow, so output is written both as a CSV 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.

4.3.2 Results Repository RDF

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[12]), 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.

[10] http://www.music-ir.org/mirex/wiki/2010:Main_Page
[11] http://jena.sourceforge.net/

For example, output from the genre classifiers is modelled using a locally declared *GenreAssociation* subclass of Association, which has as its subject a Signal instance (derived from the AudioFile via the Audio File Repository linked data), and as its object a *MusicGenre* instance as defined by the DBpedia[2] ontology.

Further Provenance Vocabulary is used to record the Meandre flow execution instance that performed the analysis (*createdBy*), the classifier within the flow (*usedGuideline*), and the AudioFile input to the analysis (*usedData*, as distinct from the parent Performance of which the AudioFile is a derivative artefact). The CSV file containing frame-by-frame analysis is linked using the Opaque Features File ontology[12].

4.4 myExperiment Workflow Management

The myExperiment[9] 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 Meandre workflows, as used by the NEMA system and the Country/Country prototype, has been added to myExperiment. This includes a preview page for Meandre flows, the same ability to share and manage Meandre flows as for Taverna, and functionality to enact the flow on a specified Meandre flow server (figure 3(ii)). The underlying implementation stores MAU files (a complete self-contained Meandre workflow including executable components and workflow metadata) within the myExperiment system.

The myExperiment API has also been extended, to support importing collections from the Country/-Country Collection Builder (section 4.2). 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 section 4.3).

A link utilising 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.

4.5 Results Viewer Web Application

The final service provided as part of the Country/Country prototype system is a web application which allows a researcher to view the analysis results, cross-reference against collections, and combine the analysis with other linked data sources. More than any other component the Results Viewer is a proof-of-concept that highlights only a select number of the many possible data sources and combinations.

The Results Viewer demonstration implementation (figure 3 (iii)) begins by combining two linked data sets: it takes a collection (as created in the Collection Builder, section 4.2) and queries the Result Repository SPARQL endpoint (section 4.3) matching Association results for Signal contained in the collection(s). The demonstrator is focussed on our country-centric genre analysis scenario: using country derived collections cross-referenced with result data a number of statistics and visualisations 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;
- For each signal in the collection: a full listing of genre weightings from each classifier, a playback page which 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

[12] http://purl.org/ontology/off/

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 link 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 data sets: not only do GeoNames (and the linked Jamendo data set and the Country/Country collections) and DBpedia use different ontologies for countries, but artists in DBpedia can 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 geographic 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[13].

5 Conclusions and Future Work

Researchers in the field of Music Information Retrieval (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 amongst 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 digitised, their entanglement with rights management will only make the curation and distribution of ever larger data sets a more complicated and time-consuming task. Even when audio data is freely available, a difficult balance must be found between the need for comparative evaluation of approaches using widely shared, understood, and re-usable data sets, and the avoidance of over-fitting an algorithm during development when a specific data set is repeatedly used for testing.

Evaluation also requires a common structure into which analytic 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[4, 1] bring the associated problem of managing ever greater quantities of result data. 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 which enables highly distributed, lightweight, and dynamic services when publishing data; 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.

Our case study has proven to be a very useful vehicle to exercise our design principles and to obtain feedback from the MIR, Linked Data and e-Research communities [17] 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 approach and technologies are more generic and widely applicable. The common data model of RDF extends a myriad of possibilities for linking with data models and categorisations within and without the MIR community, which future iterations of this architecture will address.

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

[13] http://www.bbc.co.uk/music/

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 re-use and validation by peers. While a researcher may not immediately or directly recognise the benefits of idealised Linked Data Principles, such practical benefits must surely be an attractive motivation which 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 re-use and automation grows further.

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 (the 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. Finally, this research has demonstrated how RESTful and Linked Data techniques enable the distributed serving and separation of content and metadata, and 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.

Acknowledgements

This work was carried out through the *Structural Analysis of Large Amounts of Musical Information* project funded by the JISC Digitisation and e-Content programme as a part of the international *Digging into Data challenge,* also funded by NSERC and SSHRC, and 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.

References

[1] E. Al-Shakarchi, P. Cozza, A. Harrison, C. Mastroianni, M. Shields, D. Talia, and I. Taylor. Distributing workflows over a ubiquitous p2p network. *Scientific Programming*, 15(4):269–281, 2007.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference*, 2007.

[3] Sean Bechhofer, John Ainsworth, Jiten Bhagat, Iain Buchan, Philip Couch, Don Cruickshank, David De Roure, Mark Delderfield, Ian Dunlop, Matthew Gamble, Carole Goble, Danius Michaelides, Paolo Missier, Stuart Owen, David Newman, and Shoaib Sufi. Why linked data is not enough for scientists. *eScience, IEEE International Conference on*, 0:300–307, 2010.

[4] A. Berenzweig, B. Logan, D.P.W. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[6] C. Cannam, C. Landone, M. Sandler, and J.P. Bello. The sonic visualiser: A visualisation platform for semantic descriptors from musical signals. In *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 324–327, 2006.

[7] Downie De Roure, D. SALAMI: Structural Analysis of Large Amounts of Music Information. In *UK e-Science All Hands Meeting 2010, Software Sustainbility Workshop*, September 2010.

[8] R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures.* PhD thesis, Information and Computer Science, University of California, Irvine, California, USA, 2000.

[9] C.A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, et al. myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, 2010.

[10] O. Hartig and J. Zhao. Publishing and Consuming Provenance Metadata on the Web of Linked Data. In *Proceedings of the 3rd International Provenance and Annotation Workshop (IPAW)*, June 2010.

[11] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[12] K. Kurt Jacobson, Y. Yves Raimond, and M. Sandler. An Ecosystem for Transparent Music Similarity in an Open World. In *Proceedings of the International Conference on Music Information Retrieval 2009*, pages 33–38, October 2009.

[13] C. Lagoze and H. Van de Sompel. Object Reuse and Exchange (OAI-ORE). Technical report, Open Achives Initiative, 2008.

[14] X. Llorà, B. Ács, L. Auvil, B. Capitanu, M. Welge, and D. Goldberg. Meandre: Semantic-Driven Data-Intensive Flows in the Clouds. In *4th IEEE International Conference on eScience*, pages 238–245, December 2008.

[15] C. McKay, J. A. Burgoyne, J. Thompson, and I. Fujinaga. Using ACE XML 2.0 to store and share feature, instance and class data for musical classification. In *Proceedings of the International Society for Music Information Retrieval Conference 2009*, pages 303–8, October 2009.

[16] C. McKay and I. Fujinaga. jMIR: Tools for automatic music classification. In *Proceedings of the International Computer Music Conference 2009*, pages 65–8, October 2009.

[17] Kevin R. Page, Benjamin Fields, Bart J. Nagel, Gianni O'Neill, David C. De Roure, and Tim Crawford. Semantics for music analysis through linked data: How country is my country? *eScience, IEEE International Conference on*, 0:41–48, 2010.

[18] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson. The music ontology. In *Proceedings of the International Conference on Music Information Retrieval*, pages 417–422, 2007.

[19] Y. Raimond and M. Sandler. A Web of Musical Information. In *Proceedings of the International Conference on Music Information Retrieval 2008*, pages 263–268, September 2008.

[20] L. Richardson and S. Ruby. *RESTful Web Services*. O'Reilly & Associates, May 2007.

[21] L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web. W3C Semantic Web Education and Outreach Interest Group Note, 31 March 2008.

[22] Segolene M. Tarte, David C. H. Wallom, Pin Hu, Kang Tang, and Tiejun Ma. An image processing portal and web-service for the study of ancient documents. *e-Science and Grid Computing, International Conference on*, 0:14–19, 2009.

[23] K. R. Taylor, R. J. Gledhill, J. W. Essex, J. G. Frey, S. W. Harris, and D. C. De Roure. Bringing chemical data onto the semantic web. *Journal of Chemical Information and Modeling*, 46(3):939–952, 2006.

[24] K. West, A. Kumar, A. Shirk, G. Zhu, J. Downie, A. Ehmann, and M. Bay. The Networked Environment for Music Analysis (NEMA). In *6th IEEE World Congress on Services*, pages 314–317, July 2010.

Appendix

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 location is encoded by GeoNames[14].

```
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 .
   { ?basednear geo:inCountry <http://www.geonames.org/countries/#BE> }
   OPTIONAL { ?basednear geo:inCountry ?country . }
   ?record
      a mo:Record ;
      foaf:maker ?artist ;
      mo:track ?track ;
      dc:date ?recorddate ;
      dc:title ?recordname .
   ?track
      dc:title ?trackname ;
      mo:track_number ?tracknumber .
   ?signal
      mo:published_as ?track .
}
```

[14] http://www.geonames.org/ontology/