Scaling Digital Humanities on (and utilising) the Web

Kevin Page
Oxford e-Research Centre, University of Oxford, UK
Web and Internet Science, University of Southampton, UK

Osaka Symposium on Digital Humanities, September 2011

Overview

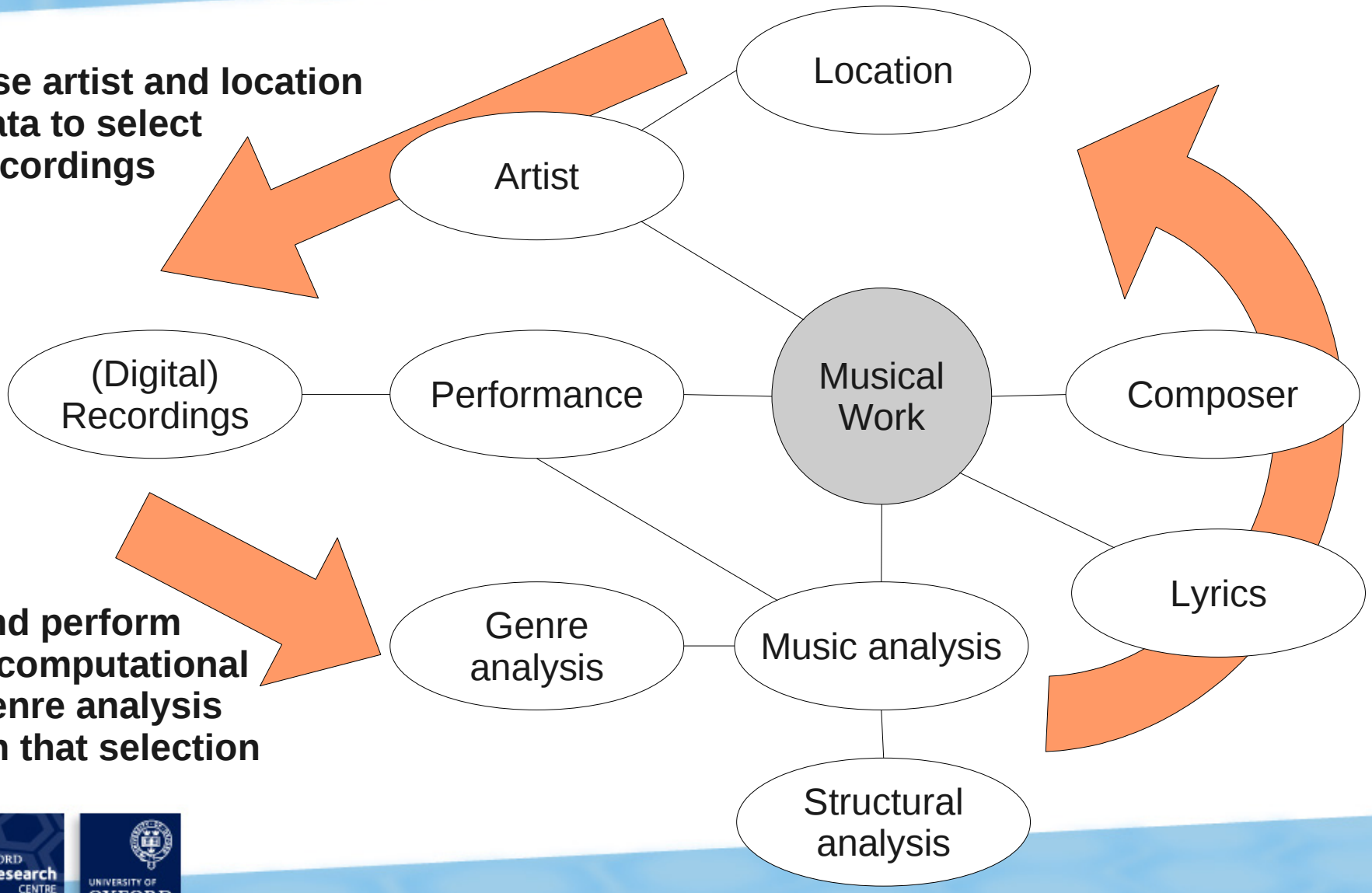
- Motivation
  - In general, and for computational musicology
- Our approach
  - Embracing Web architecture, the Semantic Web, and Linked Data
- A case study
  - How country is my country?

Motivation: in general

- When knowledge has been generated, we should capitalise on its value by
  - capturing it
  - publishing it
  - using it
  - linking it
  - re-using and building upon it (“unintentionally”?)

Motivation: a music example (simplified)

Use artist and location data to select recordings



and perform a computational genre analysis on that selection

Motivation: a music example

- Each of these conceptual areas is a specialisation
  - which might be the subject of scholarly study
  - or computational analysis
  - or crowdsourcing, etc.
- There will be overlap
  - one person's metadata is another person's data
  - we can build upon others specialisation and knowledge
- We do not expect complexity to vanish
  - but where it has been studied it should be scaled, shared, and *linked*

Our Approach

Don't just put Digital Humanities content *on* the Web...

...but use and build upon Web Architecture to scale Digital Humanities activity

The value is in the linking.

Advantages of Web Architecture

- Proven scale and distribution
  - an inbuilt mechanism for unique resource identification and addressing
- The primacy of linking
- Mechanisms to support a wide variety of content
- Easy to develop using Web Application Programming Interfaces (APIs)

From a technical perspective

- A Resource Oriented Architecture
- A Semantic Web
  - RDF: a flexible, extensible, common data model
    - not just another XML markup!
  - Ontologies: to capture and scale specialised knowledge
  - SPARQL: a common query interface
- Linked Data
  - a movement to publish *and link* RDF to create a web of data

A case study

“How country is my country?”

System architecture principles

- Multiple repositories (...datasets, viewers, applications)
- Everything (*linked*) is RDF
  - publish as linked data
  - *and make use of existing linked data*
- Be RESTful and adopt Web Architecture
- Lower barriers to using the data and developing domain applications
  - lightweight web APIs
  - encapsulate and scale complexity in ontologies

OXFORD e-Research CENTRE

UNIVERSITY OF OXFORD

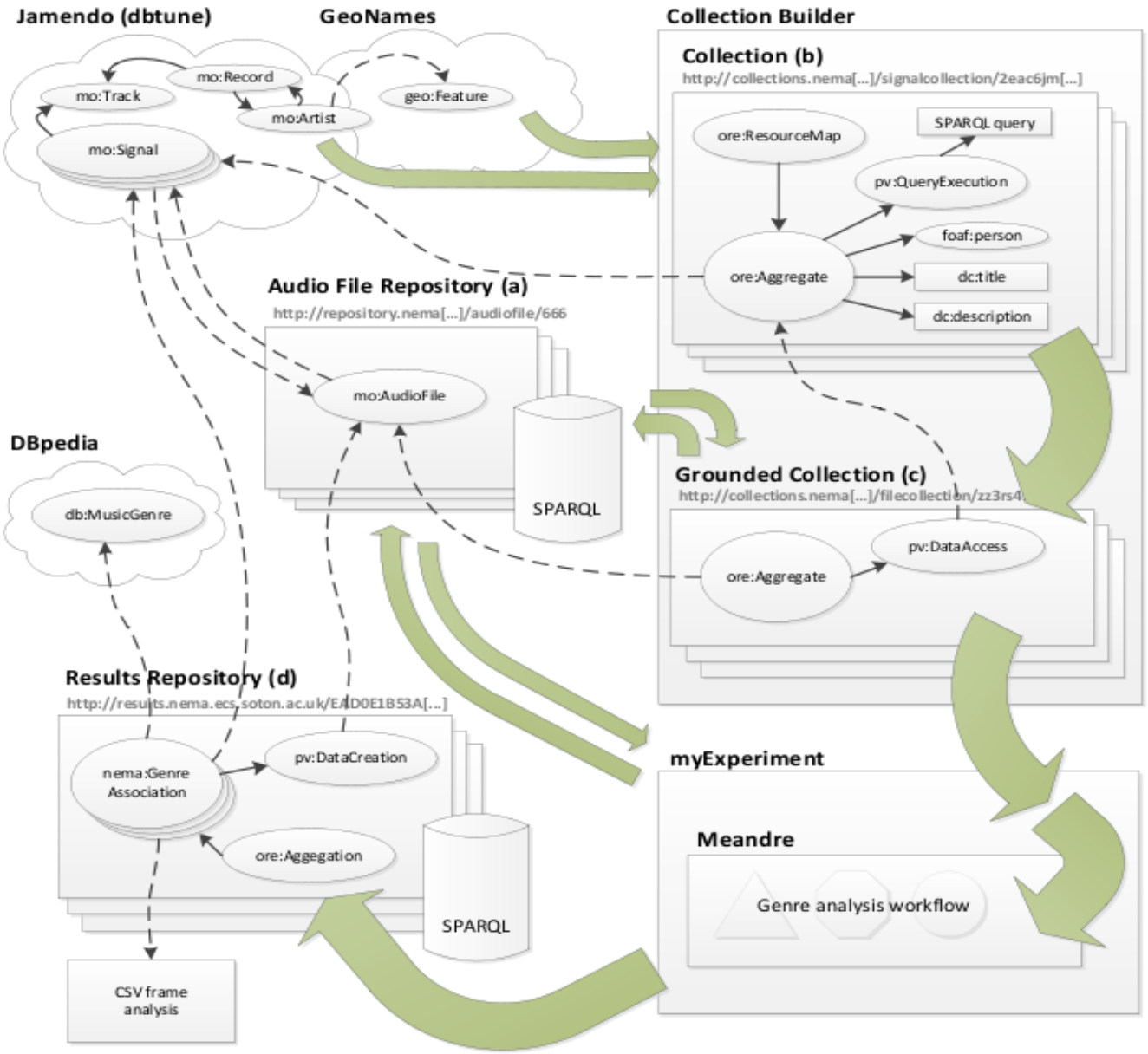
How is this manifest in the system?

- Clearly identifying, and delineating, resources
  - sometimes separating out functions previously conflated
- Serving resources as linked data using standard web services and access mechanisms (HTTP)
- Utilising appropriate – and multiple - domain and system ontologies
- Everything (linked) is RDF Linked Data
  - HTTP URIs that persist across the system (& web)
  - SPARQL provided for querying

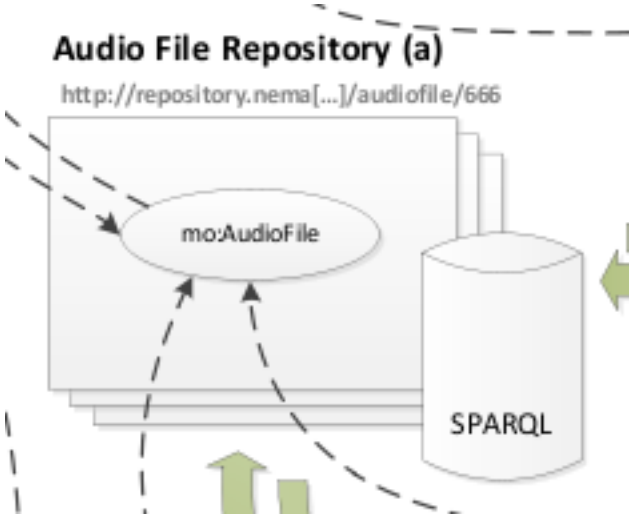
System elements

- Audio File Repositories (signal)
- Music Collections
- Algorithms and workflow
- Algorithmic output
- Results and findings

... all joined through a web of linked data


Jamendo (dbtune)
GeoNames
Collection Builder
mo:Record
mo:Track
mo:Artist
geo:Feature
mo:Signal
Collection (b)
http://collections.nema[...]/signalcollection/2eac6jm[...]
ore:ResourceMap
SPARQL query
pv:QueryExecution
foaf:person
ore:Aggregate
dc:title
dc:description
Audio File Repository (a)
http://repository.nema[...]/audiofile/666
mo:AudioFile
SPARQL
DBpedia
db:MusicGenre
Grounded Collection (c)
pv:DataAccess
ore:Aggregate
Results Repository (d)
http://results.nema.ecs.soton.ac.uk/EAD0E1B53A[...]
nema:Genre Association
pv:DataCreation
ore:Aggegation
SPARQL
CSV frame analysis
myExperiment
Meandre
Genre analysis workflow

Audio File repository



Music Information Retrieval (MIR)

- Focusses on the algorithmic extraction of information from music
- Most often a combination of *feature extraction* (signal processing) and *classification* (machine learning)
- An MIR researcher might typically:
  i. Assemble a collection of audio input (aka signal)
  ii. Apply the algorithm to the input
  iii. Publish and evaluate algorithm output

MIR systems challenges

- Exchange of music is often restricted
  - licensing and copyright
  - quantity of data
- For comparative evaluation, data sets must be
  - widely shared
  - understood
  - re-usable
- But algorithm development is susceptible to overfitting

MIR systems contexts

- MIREX
  - Music Information Retrieval Evaluation eXchange
  - Annual evaluation
  - ~20 tasks
- The SALAMI project
  - Structural Analysis of Large Amounts of Music Information
  - 350,000 songs / 23,000 hours
  - *Publication of collections, ground truth, and results as a community resource*

Existing MIR systems

- A wide variety of languages, software engineering approaches, and architectures
- Often built to solve a particular MIR problem and expanded to address others
- Systems interaction through
  - plugins
  - shared libraries
  - syntactic serialisation and file exchange
  - some semantics used, but as an enhancement to traditional systems

One trail through the system

- Audio File Repositories (the invisible groundwork)
- Create a collection of music
  - *find works by artists from a particular country*
- Find available audio files that record that music
  - *“ground” the collection*
- Pass the collection to an MIR workflow
  - *genre analysis*
- View and analysis the output
  - *how country is my country?*

Results viewer

Summary

- Linked data works
  - Web Architecture works
- Clear benefits in using URIs and ontologies
  - take advantage of existing linked data
  - publish your own linked data for others to take advantage of
  - and improve the link sparsity in the (semantic) web
- Modifications to software and systems are required
  - but they are not a huge burden
  - complexity is condensed into ontologies
  - bespoke application development is simplified
  - the Semantic Web browser is a web browser

Acknowledgements

Authors

David De Roure
University of Oxford
Ben Fields, Tim Crawford
Goldsmiths University of London
J. Stephen Downie
University of Illinois
Ichiro Fujinaga
McGill University

Funding

Andrew W. Mellon Foundation (NEMA)
JISC (SALAMI)

http://www.nema.linkedmusic.org/