Authority and Judgement in the Digital Archive

Alan Dix
University of Birmingham,
Birmingham, B15 2TT, UK
and Talis, 48 Frederick Street,
Birmingham B1 3HN, UK
alan@hcibook.com

Rachel Cowgill
University of Huddersfield
Queensgate,
Huddersfield, HD1 3DH, UK
r.e.cowgill@hud.ac.uk

Christina Bashford
University of Illinois at Urbana-Champaign
1114 W. Nevada Street
Urbana, IL 61801, USA
bashford@illinois.edu

Simon McVeigh
Goldsmiths, University of London
New Cross
London SE14 6NW, UK
S.McVeigh@gold.ac.uk

Rupert Ridgewell
British Library
96 Euston Road
London, NW1 2DB, UK
Rupert.Ridgewell@bl.uk

http://inconcert.datatodata.com

ABSTRACT

The transformative promise of the digital humanities is not without problems. This paper looks at digital archive curation using a database of 19th-century London concerts as a case study. We examine some of the barriers faced in its development, related to expertise, volume and complexity, the gap between cost and benefit, and the desire for an authoritative and complete dataset that forces a particular linear process of curation. We explore the potential for more radical approaches where curation and use are interleaved, and where digitally maintained provenance allows professional judgement to be applied to incomplete, crowdsourced, or automatically processed data.

Categories and Subject Descriptors

Applied Computing – *performing arts, digital libraries and archives*; Information Systems – *data provenance*; **Human-Centered Computing** – *interaction design*

General Terms

Design, Human Factors.

Keywords

Musicology, performance history, digital archives, digital humanities, ephemera, concerts, linked data, open data.

1. INTRODUCTION

In this paper we examine the potential of digital technologies to transform the nature of humanities archives. We argue that this potential can only be fully realised if the design takes professional practices and academic values into account, and furthermore that technology can transform the very processes through which we conduct research if designed in ways that are sympathetic to and preserve the deeper academic values they represent. As the focus of our discussion we take a case study from musicology – a database of 19th-century London concerts.

In musicology, the emergence of digital technologies coincided with the experience of a 'cultural turn' – a broadening of disciplinary focus from the age-old preoccupation with producers of musical works (composers) to include the role of performers and musical consumers (audiences, critics, institutions, taste-makers) in shaping musical culture. Such a shift in perspective required the development of new methodologies to explore previously neglected print sources – newspaper advertisements and criticism, for example, or ephemera such as concert programmes – and technologies were applied to make these newly accessible through pioneering digital library initiatives.

The research group *In Concert: Towards a Collaborative Digital Archive of Musical Ephemera* [15] is investigating current and future standards in the development, curation, and use of data in this rapidly growing area of contemporary humanities scholarship.

The next section examines some background to this project, the more general issues of archives in the digital humanities, and the specific development of the *Concert Life in 19th-Century London* database (*CL19*). This dataset is then used, in section 3, to inform analysis of the broad processes of data and scholarship involved in digital archive production and uses. Section 4 looks at barriers to the full exploitation of the *CL19* database and how these relate to the fundamental values and motivations of musicologists. This leads, in section 5, to a re-examination of the way digital technology could support patterns of work that respect the underlying professional values of the academic, but are radically different from existing practice. Finally, in section 6, we describe some of the practical work we are currently engaged in as part of the *In Concert* project, based on the analysis in this paper.

2. BACKGROUND

2.1 Digital Humanities Transformations

Humanities researchers have been quick to explore the potential of digital technologies for gathering, systematising, and querying rich bodies of data. Sometimes this has simply made existing practices more efficient, for example, using a digital search instead of a manual catalogue. Sometimes it has opened up new possibilities through computational capabilities, for example, in the statistical analysis of corpora for authorship fingerprints.

In addition, the sheer volume of data that can be managed by digital technology changes the kinds of material that can be dealt with effectively, leading to the potential to store and examine the vast volume of minor ephemera of day-to-day life as well as the few 'special' outputs of artistic luminaries. This 'infinite archive' [3] means that preservation, storage, and indexing become less problematic, but now the barrier becomes one of sifting the interesting from the mundane and the exceptional from the typical, lest the data become, like Borges' map, coincident with the world itself [6].

So there are clear transformative effects at a technical level; however, the emergence of digital technology in the humanities is also having transformative effects on traditional academic practices, sometimes simply expanding their reach, but also sometimes radically challenging their essence. In philology, for example, advice on grammatical and typographic conventions for ancient languages has traditionally been built from years of individual experience. Digital methods now mean it is possible to validate these conventions statistically, sometimes leading to conflicts between 'proper' forms (according to traditional academe) and actual use (based on data processing); rather like the way 'their' defies rules for spelling.

In short, digital technology is leading to a re-examination of the practices as well as products of humanities research. This may be resisted, as it may challenge existing structures of academic authority (as in philology), but it can be seen as an opportunity to dig beneath the surface of practices as they have evolved in relation to traditional media and methods, to reveal underlying values and principles that can be re-embodied in digital practice.

2.2 Concert Life in 19th-Century London

More than ten years ago, Cowgill, Bashford, and McVeigh created a database of 19th-century London concerts from material relating to concert performances published in newspaper and magazine sources. These comprised a mixture of announcements, advertisements, puffs, and reviews, but also had the capacity to include other ephemeral sources, such as concert programmes [2]. The initial capture was by a team of research assistants, as part of a project partly funded by the AHRB, creating a structured database from the source documents.[1]

The structure of the database was informed by an earlier database of 18th-century London concerts compiled in the 1980s by McVeigh from all available newspaper sources – *Calendar of London Concerts 1750–1800* [17]. This was a very early example of the use of digital technology in musicology and has been used extensively as a research resource although the technology is now very dated. While the volume of 18th-century concerts was substantial (4,000 records), the increase in popularity and frequency of concerts in the 19th century made the exhaustive collation of all related newspaper material impossible; instead, representative years at 20-year intervals between 1815 and 1895 were used in the *CL19* database – a methodology known as 'slice history', which had been pioneered by social historians.

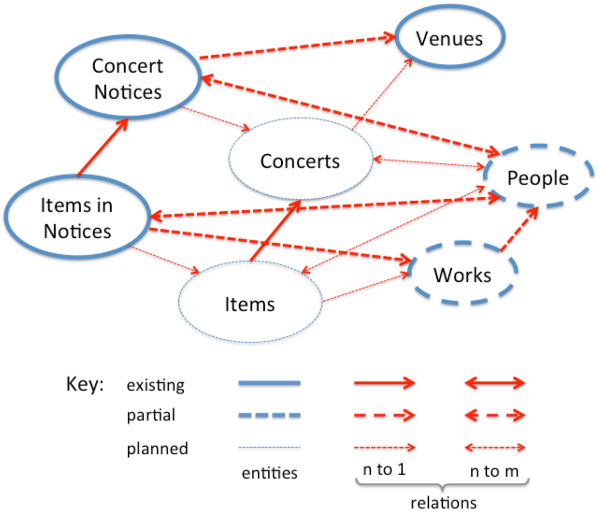

Figure 1. Structure of the *Concert Life in 19th-Century London* database project (*CL19*)

The result was a database of information about concerts, in which one record was created for each notice relating to a concert that was published in each of the selected sources, and one or more additional records corresponding to each item of information given in the source about the concert itself (see fig. 1).

A single concert might be documented in several forms, each with partial information, which when considered together give more complete information. So there needs to be a task of record linkage [9] (which was programmed into the original *CL19* database structure) comprising the identification of notices that correspond to the same concert, and then populating the final concert record based on the partial information in each. In addition, concert-related notices in newspapers mention venues, works, composers, performers, and organisers, in often inconsistent forms, which need to be lined up with corresponding authoritative records for each place, work, or person.

This in itself has proved a valuable research resource for the academic partners; however, two additional phases of (a) data curation and (b) analysis were also planned. These two phases have not been completed due to a combination of technological and professional constraints. A particular problem has been that a large expert time commitment was required to finish the curation process before musicological questions concerning the development of 19th-century performance culture could be investigated. This was exacerbated by the relational database technology available at the time.

An important aim of the *In Concert* project is to break this bottleneck, in particular by relaxing the constraint for a complete authoritative curation (data-cleaning, validation, and matching) before musicological analysis and enquiry can be started.

3. PROCESSES AND METHODS

In order to understand barriers to progress for digital archive creation in general, we first need to understand the activities and methods used in collection, curation, and use.

[1] *Concert Life in 19th-Century London* database project, funded by the University of Huddersfield and Oxford Brookes University (1997–2001), and the Arts and Humanities Research Board (UK) and University of Leeds (2001–04). Research Assistants: AHRB Research Assistant (with studentship) Ann Royle (1999–2004). Project Research Assistants: Stephen Follows (2001–03); Vania Schittenhelm (2000–02); Emma Peake (2000); Michelle Brachet (1998–2000); Clare Lyon (1998–2000); Jamie Savan (1998–99); James Marshall (1998–99). We would like to record our thanks to the research assistants for their integral contribution to Phase I of the project.

3.1 Raw Source to Authoritative Data

The *CL19* database project was conceived to follow what is probably a classic data-production life cycle in the digital humanities. The investigative framework having been established, a corpus of primary sources was selected and digitised though several stages of interpretation to create a definitive database, which could then be used in the production of quality academic research (fig. 2).

Semi-expert researchers were deployed in the early stage, but more expertise was needed in later stages as interpretation became more complex and required greater academic judgement to deal with inconsistencies in and incompleteness of the primary data.

The process was driven by a particular set of academic concerns – the selection, codification, and interrogation of a highly structured body of material relating to 19th-century concerts, rather than, say, murder reports or accounts of scientific meetings – and is, in a broad sense, goal directed. Within these general concerns, however, it aimed to create a broadly useful resource that was open both for statistical questions or more detailed analysis of specific trends and aspects, such as the development of individual concert series, venues, repertoires, programming practices, or artists.

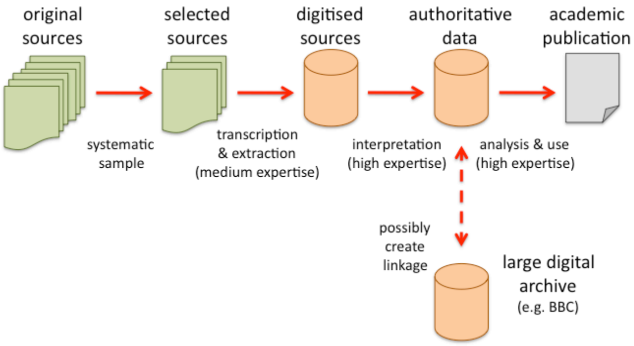
original sources
selected sources
digitised sources
authoritative data
academic publication
systematic sample
transcription & extraction (medium expertise)
interpretation (high expertise)
analysis & use (high expertise)
possibly create linkage
large digital archive (e.g. BBC)

Figure 2. *Concert Life in 19th-Century London* database – a classic digital humanities process?

3.2 Digital Archives

Digital archive production is similar to the process outlined here, except that the focus is broader in terms of ultimate purpose, i.e. to create a resource by making available as wide a selection of materials to as wide an audience as possible (fig. 3).

Both the more goal-directed *CL19* database and more broadly conceived digital archives share a crucial aspect, however, which is the production of a complete (although possibly selected), unbiased, and authoritative resource for further analysis.

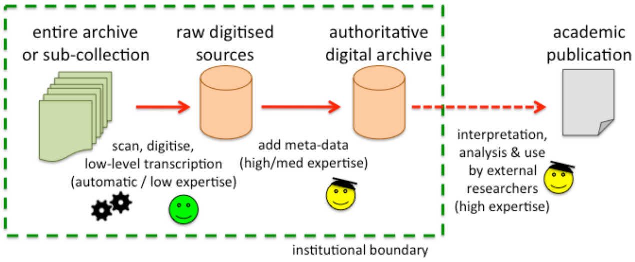
entire archive or sub-collection
raw digitised sources
authoritative digital archive
academic publication
scan, digitise, low-level transcription (automatic / low expertise)
add meta-data (high/med expertise)
interpretation, analysis & use by external researchers (high expertise)
institutional boundary

Figure 3. Digital archive production and use

Some early stages of the digital archiving process, notably scanning and transcription, require little professional expertise, although some, such as art-work photography, require specific media skills. The raw data still need meta-data in order to be useful, which may require semi-professional input; however, the aim is to produce accessible and indexed versions of raw resources for use by external researchers, who are expected to add the final interpretative analysis.

3.3 Crowdsourcing and Digital Archives

Many large collections are adopting crowdsourcing in digitisation projects to augment efforts in some of the less expert, but hard to automate stages. This may be at a very basic level, such as the use of CAPTCHA codes to supplement OCR of the *New York Times* archive and Google Books [1], or may require generic human expertise, such as the matching of geographical features on antique maps with their appearance on modern maps of the same area [10].

Crowdsourcing initiatives typically have to add additional mechanisms to ensure the quality of the resulting data. These may include multiple entry of the same data by different encoders, or cross-checking the validity of values (some things are hard to create automatically, but easier to check).

Crowdsourcing is similar to a conventional digitisation exercise, in so far as the final use of the data is typically by researchers from outside of the institution housing the digital archive (fig. 4). However, there are differences. For OCR or similar processes, academic users will be aware of the nature of the digitisation process, and so be alert to any potential problems, such as the unintended 'correction' of archaic spellings. In contrast, with a partially crowdsourced project, academics are more likely to seek corroboration of data.

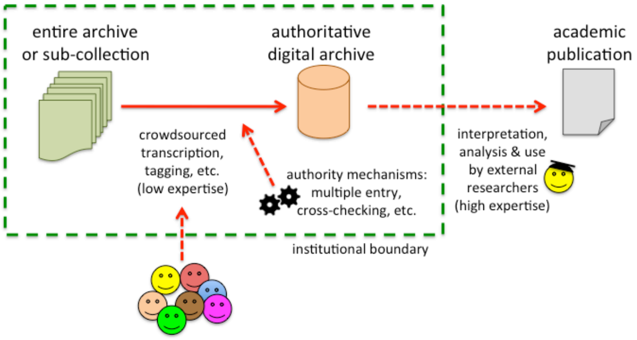
entire archive or sub-collection
authoritive digital archive
academic publication
crowdsourced transcription, tagging, etc. (low expertise)
authority mechanisms: multiple entry, cross-checking, etc.
interpretation, analysis & use by external researchers (high expertise)
institutional boundary

Figure 4. Crowdsourced digital archives

3.4 Google: Automation and Volume

It is interesting to contrast the principles outlined above with the Google approach. Basically, Google focuses first on low-hanging fruits – that is, large data sources that can easily be digitised, either automatically or with minimal expertise (for example, by driving round the streets!). This preliminary work is then supplemented by massive automatic analysis. Although the data sources are of variable quality, the sheer quantity has proved 'unreasonably effective' in dealing with many kinds of linguistic and related questions [13]. However, there are interesting areas, such as digital mapping, where human corrections are needed.

In some ways research using 'big data' can be regarded as objective, taking all the available data and often applying massive-scale machine learning. However, this carries its own biases as Internet knowledge is not all knowledge and in particular tends to over-represent western countries, professional classes, and younger people. Google has attempted to redress some of this bias in its Indigenous Mapping project [12], although even this has been critiqued as it is still within the overall structures of Google [11].

While the application area is quite far from musicology, it is a reminder that no data and no archive, digital or otherwise, is without some level of selection bias. Rather than attempting to completely remove such bias (an impossible task), we can instead attempt to make clear the sources and their respective selection criteria, biases, and viewpoints.

3.5 Linked and Open Data

Linked data [5,14] is part of the realisation of Berners Lee's goal of a Semantic Web of data [4], readable by computer in parallel with the web of documents understandable by people. The idea is to connect together disparate data sources using URIs as a *lingua franca* to represent individual entities (people, documents, or pieces of music).

The potential for the use of linked data and semantic web technologies in music was recognised early, allowing composers, performers, genres, and other core concepts to be connected. In particular, the BBC developed a Music dataset early in its explorations of semantic web technology [19], and it was one of the earliest exemplars used in the mSpace browser [22]. More recently shared ontologies and linked data are to underlie the API of the *Transforming Musicology* project [24].

4. BARRIERS TO PROGRESS

Having understood the process and also potential alternative technologies available, we will now look at why the second phase of the *CL19* database stalled, and barriers to extending the work.

4.1 Challenges at Phase II

As described earlier, the *CL19* database contains raw data from print sources sampled at specific years during the 19th century, but this was conceived as the first stage of a larger plan to apply expertise and judgement to this corpus to create as comprehensive and authoritative a database of concerts as possible. To gain a head start with this, the research assistants incorporated into their transcriptions possible identifications or other messages to the core team (according to strict protocols) with the intention that this would form a secure basis for future work.

Unfortunately the larger work stalled, partly for technical and professional reasons (investigators moving jobs, changing institutional priorities, inaccessibility of the data), but partly because of the sheer volume of high-expertise work needed *before* any analysis could be done to create publishable results.

Whenever there is a large gap between effort and reward, systems stall or fail; this is particularly the case in academia where issues of tenure, funding, or research assessment create a focus on relatively short-term outputs.

In short the problems are:

- high expertise needed
- volume, complexity, and inconsistency of material
- gap between cost and benefit

This is unlikely to be a unique situation for the digital humanities.

4.2 Open Data, Linked Data, and Crowdsourcing

Early discussions between the technologist and musicologists on the *In Concert* team included various ways in which the existing data could be made more valuable. Potential ideas were proposed corresponding to the technologies described in the previous section:

- Making the data openly available on the web, so that other researchers can use it.
- Linking the data to third-party sources, for example, linking composers to the DBpedia information (a datasource extracted from *Wikipedia*) [8].
- Using crowdsourcing to complete some of the further steps required, perhaps recruiting knowledgeable amateurs to work on specific tasks under carefully controlled conditions (rather in the manner of the original research assistants) or simply accepting whatever data and interpretation might be contributed by visitors to and users of the resource.

Despite a sympathetic hearing, it is probably fair to say that initially none of these suggestions were greeted warmly by the musicologists – although linking to sources such as British Library records was seen as more acceptable than to a dataset with *Wikipedia* associations, and automatic processing was seen as more acceptable than gathering crowdsourced material.

4.3 Authoritative and Complete

Two key concerns for digital archives in general, and the *CL19* database in particular, are the desire to be:

- ***authoritative*** and of known quality, so that the data can be used reliably for further interpretation
- ***complete***, or at least sampled in a well-controlled and well-documented manner, so that bias in any trends observed or statistical analysis derived from the data is minimised

Some of the distrust of crowdsourcing is due to worries about the quality of results; the work delegated to non-experts has to be carefully chosen, so that it does not require deep domain knowledge and can be verified automatically. Even in relatively simple tasks, such as word transcription, there may be a tendency, for example, to 'correct' archaic or variant spellings, to expand abbreviations, or to compensate for the frequent flaws in printing by supplying missing letters based on a combination of deduction and guesswork.

Of course no result is entirely authoritative; even the results of well-respected academics are understood to be influenced by particular perspectives, contexts, and styles of working, which are taken into account when other researchers critique and build on their work. Even 'original' sources may be the result (in older documents) of copying or printing processes, which have to be dealt with. For example, Biblical scholars faced with discrepancies between sources – one with an oddly phrased passage and another where it makes easier sense – now tend to assume the former is more likely to be the original reading and the latter a version that has been 'corrected', maybe even unconsciously, during the process of transcription.

Likewise, even if one has a complete or unbiased sample of an archive, the archive itself is subject to historical bias determined by collection policies and the differing chances of survival for documents of varying types and condition. In the case of the *Concert Life in 19th-Century London* database, one could sample years systematically by extracting all material relating to concerts published in newspapers during that period; however, this would only be a complete sample of concerts publicised in newspapers, which may, therefore, not include musical performances taking place before an audience but in a less 'public' setting. It is a matter of academic judgement as to whether coverage is deemed sufficient to answer a specific question, and how to nuance the findings and deal with any biases introduced. Such issues were discussed at length during the early planning of the *CL19* project and design of the database: to address these limitations some data was captured from concert programmes, advertisements triangulated with reviews, and so on.

Scholars in history and prehistory have similarly had to develop methodologies to work around a recording bias towards 'kings and queens' and 'stones and bones' respectively, knowing that the extant record does not preserve uniformly. Archives themselves are shaped by bias at a structural level, having been collected and preserved for particular purposes, often by individuals or institutions with particular priorities.

Crucially, the scholar brings to bear academic judgement in individual scenarios to answer specific questions based on an assessment of the provenance, authority, quality, coverage, and bias of resources.

4.4 Digital Acontextuality

Issues of authority are complicated further by the difficulty of assessing the reliability of digital records.

A traditional scholar would need to travel to the physical collection in order to consult an archive. Its location – the half-forgotten attic of an old house, or a modern library – would itself tell something of the origins and reliability of the material. The index might be clearly printed on 3x5" cards, roughly handwritten in an old shoe box, or be non-existent, leaving the researcher to leaf through unsorted papers without guidance. The source might be a well-printed book or memories scrawled quickly in a notebook preserved by chance. The physical form, location, and disposition of the artefact tell as much as the words written on it [20].

Now, the modern scholar enters search terms online and receives uniformly printed results garnered from a variety of archives, themselves drawing on a variety of original sources. The source might be identified in a field, but myriad clues implicit in the original are undetectable when the researcher is one step removed. Digital materials often rob the raw information of its context.

This is rather like Socrates' concerns about the written word itself; that it is flat, and cannot be interrogated [18]. While you cannot ask explicit questions of a printed manuscript or newspaper cutting, the physical form can give many answers.

In principle, the digital archive should be more perspicuous, but the reality is often the opposite.

4.5 Openness: Reward and Control

These issues of authority also partially explain the *CL19* partners' resistance to publishing the Phase I data; editorial annotations in the dataset were intended for use within the team and could be confusing to a third party. These annotations were intended to bootstrap the interpretation in Phase II, and this highlights further issues that can lead to resistance to publishing data, connected not so much to the professional values of authority, but to the academic reward system.

In general, academics get recognition for research when it is released in the form of a publication such as a journal paper or book. There are fields, for example linguistics, where dictionaries or other corpora are valued, but more broadly the publication of collated data has at best partial recognition as a valuable output. For example, while the UK Research Excellence Framework has recently recognised a 'database' as a legitimate research output in the humanities [21], the Leverhulme Trust, which supports cross-disciplinary research, explicitly rules out database creation as a principal goal for research in its guidelines for project grants:

> "*The Trust will not fund applications in which the balance between assembling a data bank or database and the related subsequent research is heavily inclined to the former*." [16]

Surprisingly, the situation is worse in computer science, where datasets are valued as 'service' to the research community, but not highly regarded as evidence of individual research.

Collecting and creating a dataset requires substantial effort, but if the rewards are only accrued by the subsequent analysis, then this creates substantial barriers to the creation of open datasets. The danger for the original researcher is that someone else gets the all-important credit on the back of the original hard and time-consuming labour.

Other issues can also cause barriers to openness. In India there are approaching one million ancient texts in archives scattered across the country. Some of these are being digitised, and the scans are made available on request to bona fide researchers. There is resistance to making them openly available (even under, say, a non-commercial licence) for fear that they will be 'misused'.

This worry appears to be partly about intellectual property issues, in case, for example, traditional medical cures lead to the development of patented drugs. This objection appears to stem partly from a misunderstanding of patent law, but the core worry seems a more deep-seated discomfort about losing control, exacerbated by the fact that many of these are also religious texts.

While most humanities archives do not share these religious and cultural sensitivities, some of this same sense of uncertainty may well be present. Releasing data openly on the web means relinquishing control, and typically having no knowledge of how that data is used, with the possibility that it might be misinterpreted, recontextualised inappropriately, or even misrepresented altogether.

5. RE-ENVISIONING DIGITAL ARCHIVE CURATION AND USE

A key question emerges from the above discussion: can we break the barriers and formulate a different way of approaching the development of digital archives?

We have already said that the central aspect of archival methodologies in the humanities is that:

> *"the scholar brings to bear academic judgement to answer specific questions based on an assessment of the provenance, authority, quality, coverage, and bias of resources."*

We take this as a lynch pin, addressing the core barriers, whilst keeping the role of academic judgement as central, but re-applying it in new ways to maximise the potential of digital technology.

We also look, for comparison and continuity, to the way in which pre-digital research was carried out with equal rigour, but using now comparatively low-tech tools such as the photocopier, notebook, and highlighter. The aim is not to emulate these tools and practices in facsimile, but, by understanding them, to know better how to design new digital tools and practices.

5.1 Expertise

Automatic methods or crowdsourcing can reduce the level of expertise required, but at the cost of a potential loss of quality and authority. The challenge here seems to be to ensure that the provenance of derived data is made apparent, so that academic judgement can be brought to bear.

This may be done on a case-by-case basis, especially for automatic analysis – for example, a count of concerts containing the keyword 'Handel' in the composer field in different years would give a good indication of the popularity of George Frideric Handel, whereas a search for 'Strauss' would be more problematic, given the number of musicians bearing that name, many from the same family.

There is something in the nature of musical data, particularly the sources that identify pieces of music – Symphony no.6, Sixth Symphony, Symphony in F, Sinfonia no. 6, Pastoral Symphony can all refer to the same Beethoven work – that makes the need for authoritative identifiers so crucial in digital musicology and, indeed, a requirement before higher-level interpretative work can be done. Musical data, it seems clear, carries a complexity way beyond that involved in the identification of a novel or painting, the majority of which would have unique titles.

Automatic analysis can also be used to offer suggestions to aid hand analysis. For example, names of composers can be matched against existing data entries in the database or external sources, such as *Grove Online* or the BBC Proms archive, and the best matches offered as options; this way, the binding of text to verified entries is controlled by an authoritative process, but made more time-efficient. But perhaps automatic analysis could be taken a step further in this case, to identify an individual or work from the context in which it appears – 'Strauss' may be preceded by 'Richard' or 'R', as opposed to 'Johann' or 'J', or, indeed, 'E', but if only 'Herr Strauss' is given in the source and a work title follows close after, then 'Richard Strauss' might be identified from a proximity to '*Don Juan*' or 'Johann Strauss, Jr' from association with the '*Blue Danube Waltz*'.

5.2 Completeness

In software-engineering systems, the linear process in Figure 2 would be described as a *waterfall model*, where each stage must finish before the next can be initiated. In contrast, *agile methods* in software production are often focused around 'use cases', that is, doing only sufficient work on a system to obtain a particular piece of functionality [23].

One may be able to answer certain questions based on only partial processing of the complete dataset. For example, in the *CL19* database there are typically multiple advertisements and reviews for the same concert; these need to be matched up and then the data from individual sources merged to give a definitive record for each concert and, where such information is given in the sources, for each item on the programme. However, there may be questions that can be answered on the basis of the matching stage alone. Simply sorting source entries by date and then ticking those that are to be linked is a relatively fast process. Many questions may require more definitive data, but various statistical queries could be answered solely from this activity.

Alternatively, one may be able to process a subset of the data entries more fully than the complete dataset. For example, if one were interested in the relative popularity of various items by Johann Strauss the younger and Johann Strauss the elder, one might select all source items containing 'Strauss' and then hand code which particular Strauss is the composer for each item performed in each concert. Over time, addressing specific questions on different occasions, according to the varying priorities of a range of researchers, the data would eventually become better coded.

5.3 Cost-Benefit Gap

Taking a more goal- or question-oriented approach to the processing of digital data also closes the cost-benefit gap, meaning a smaller investment is needed in order to obtain useful results. This more ad hoc approach does not preclude a more complete approach to interpretative analysis. Indeed, if data have been partially processed due to more goal-specific analysis, then at some point the additional work required to fill the gaps in the definitive database may become manageable. Some research questions may need to be deferred until this point.

5.4 Workflow Management

One of the benefits of waterfall approaches is ease of management. It is easy to see what has been completed and what needs to be done next, so to organise a project based on clear milestones. More ad hoc analyses will leave different portions of the database with different levels of processing.

This makes it important to have some form of workflow support, making clear what work has been done, and what remains to be completed, both to address specific goals and considering the aims and scope of the entire dataset.

Such considerations are important for helping to allocate time to data-curation activities, but also for making visible whether the data are suitable as the basis for answering particular questions. For example, if concerts have been fully coded for all entries mentioning 'Strauss' and 'Handel' in the composer field, then this subset may be usable for certain kinds of question, just as one may be able to answer certain questions by consulting an archive focused on a particular artist, institution, or period. What is crucial is that the scholar is aware of the selection criteria implicit or explicit in the data, in order to be able to exercise academic judgement in ascertaining their value for addressing a specific question.

5.5 Openness and Data Publication

A more incremental approach to data curation backed by clear workflow management means that, at an early stage, some parts of the archive have been:

- validated sufficiently to be acceptable as authoritative work
- analysed and interpreted sufficiently to allow publications based on the data

This effectively removes the major barriers to publication of the data in some form of open access. While the whole dataset may remain under embargo, those parts that are deemed suitably complete and have already yielded value to the researchers can also be made available to others.

6. DISCUSSION AND ONGOING WORK

While we have developed a relatively detailed analysis and concrete proposals, we are aware that these are still tentative and require further work to determine both effectiveness and generality. Within the *In Concert* project we are using a case-study approach to address this in both depth and breadth.

6.1 Depth: Horizontal and Vertical Slices

To deal with depth we are choosing representative 'questions', then generating bespoke solutions for each question. The aim is then to generalise from these to look at generic data-management techniques that satisfy the principles outlined in the previous section.

An example question is:

> *"How do the sites for musical performance in London change over time in terms of geographical distribution and audience demographic? (e.g. West End vs City, inner city vs developing suburbs, North vs South London, proximity to railway stations and developing tube networks)"*

This question requires some detailed work on concert venues, but there are relatively few of these. Furthermore, we can proceed by linking the venues to other datasets: notably, the *Concert Programmes* database [7], which includes a list of places that are currently being geotagged in a separate project.

More substantially, in order to obtain detailed statistics of changes and trends:

(i) Each concert needs to be matched to its venue.

(ii) Where multiple entries exist for the same concert, these need to be identified with one another.

Happily the first of these is straightforward. There are relatively few venues and they are generally quite distinct, so the majority have been automatically matched by name with a high degree of confidence. These may need to be verified by hand at some point; for initial analyses at least, this is likely to be sufficient.

The second task can also be performed automatically, by associating concerts with the same data and venue. This is more sensitive to mistakes in the original data ('11 Sept' printed as '12 Sept' is a more likely error than 'Wigmore Hall' printed as 'Albert Hall'), and critical for statistics, so a manual identification stage is more important. However, the ability simply to confirm should make this work relatively rapid, if a little tedious.

Note that this exemplifies one of two different ways in which we are able to address questions without a 'complete' dataset. Here we have what can be thought of as a *horizontal* slice through the dataset, doing a small amount of work to every part of the dataset. By contrast, a few years ago Bashford performed all stages of analysis and interpretation of a *vertical* slice – a selection of concerts that took place at a particular venue, the Wigmore Hall, in one season, 1906–07.

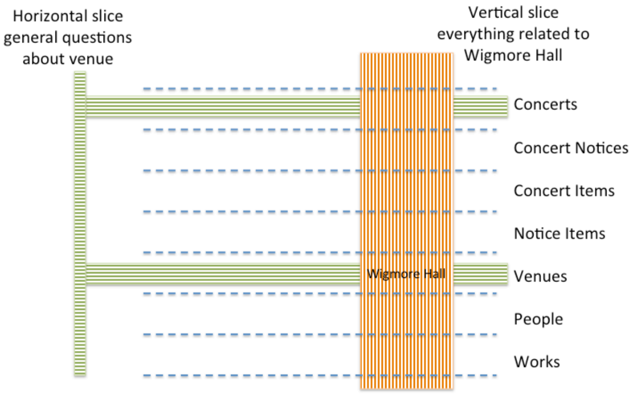

Figure 5. Partial completeness: horizontal and vertical slices

6.2 Linking Concert Data

To expand with breadth, we are connecting a number of datasets related to concert ephemera. Three of these have already been mentioned, the *Calendar of London Concerts 1750–1800, CL19*, and *Concert Programmes* databases.

The first is complete, both in the sense that it covers as fully as possible all concerts during the period 1750–1800, and in that it has been fully interpreted, identifying individual venues, people, and works, as well as identifying individual concerts from multiple sources. The *CL19* dataset, as noted, is a sample of years, and at present is only partially interpreted.

The *Concert Programmes* database is highly curated, but is a meta-dataset, describing archives and their content only, so not capturing the data contained within them. However, its list of venues is very well maintained, and can be linked to enrich the other datasets, as noted in the previous example.

Finally, we have access to scanned portions of the *Konzertprogramm Austausch* ('Concert Programme Exchange'). This series of publications was promoted by the Leipzig publisher Breitkopf & Härtel between 1894 and 1944, in order to share printed concert programmes among musical organisations in Europe, Scandinavia, Russia, and the Americas. This is an example of a dataset at the very earliest stages of curation, and will enable us to examine the extent to which useful questions can be addressed using plain OCR of these programmes, and to explore the potential for crowdsourcing the enormous task of extracting more detailed data from the scans.

In addition to interlinking these datasets, we intend to connect them to external datasets. As a preliminary step we have trialled automatic linkage to MusicBrainz, a crowdsourced web-based music resource. While the data in MusicBrainz may not be regarded as definitive, they may be good enough for certain purposes, and moreover their identifiers are used by the BBC for its music-related linked data.

We hope that these datasets of different kinds and at different stage of curation will help us to refine our understanding.

7. CONCLUSIONS

The quest to maximise the potential of digital technologies calls for a critique of traditional understandings of authority and academic judgement. Future researchers will need to be open to alternative, contingent ways of proceeding with enquiries based on incomplete and/or partially verified data.

This is not about loss of control, but about digital systems rigorously documenting varying levels of completeness and then visualising this appropriately. This will enable researchers to apply their professional judgement, factoring this incompleteness into their working methods, and will allow them to analyse general trends with a quantifiable, relative level of certainty.

It will also enable an incremental approach where the work done by these researchers in addressing specific questions and concerns is fed back into the digital system, increasing the reliability for others. The traditional model of the lone humanities scholar with absolute control over his/her materials is giving way to a collaborative or distributed model of research.

We are working towards a large-scale meta-project, involving inter-linked datasets and facilitating individual contributions in pursuit of individual projects, that all lead (no matter how indirectly) to the production of a collaborative research resource greater than what might be achieved by even a substantial body of scholars working on a single collaborative project. This will act as an exemplar of the potential for open and linked data in future humanities research.

In summary, digital technology has great potential to transform the humanities, but this will only be fully realised if digital systems are sympathetic with fundamental academic values. This will involve a reimagining of the professional processes of the humanities, however, while staying true to those underlying values. Happily, digital technology can aid this radical transformation.

8. ACKNOWLEDGEMENTS

In Concert is a mini-project of *Transforming Musicology* (http://www.transforming-musicology.org), itself funded under the AHRC Digital Transformations in the Arts and Humanities scheme.

9. REFERENCES

[1] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 12 September 2008,. 321(5895):1465-1468. DOI: 10.1126/science.1160379

[2] Bashford, C., Cowgill, R. and McVeigh, S. (2000). The Concert Life in Nineteenth-Century London Database, in *Nineteenth-Century British Music Studies*, 2, ed. by J. Dibble and B. Zon (Aldershot: Ashgate, 2000), 1-12.

[3] Bell, D.(2004). Infinite Archives, *SubStance*, Vol. 33, No. 3, Issue 105, pp. 148-161, University of Wisconsin Press. http://www.jstor.org/stable/3685549

[4] Berners-Lee, T., Hendler, J., and Lassilia, O. (2001). The semantic web. Scientific American, 284(5):34–44. DOI: 10.1038/scientificamerican0501-34

[5] Bizer, C., Heath, T., and Berners-Lee, T (2009). Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22. DOI: 10.4018/jswis.2009081901

[6] Borges, J. (1946). Del rigor en la ciencia. (tr. 'On Exactitude in Science') *Los Anales de Buenos Aires* 1.3 (Mar. 1946):53.

[7] Concert Programmes online database. accessed 25/6/2014. http://www.concertprogrammes.org.uk/html/about

[8] *DBpedia* . accessed 23/6/2014. http://dbpedia.org/About

[9] Elmagarmid, A., Ipeirotis, P., and Verykios, V. (2007). Duplicate Record Detection: A Survey, *IEEE Transactions on Knowledge and Data Engineering* 19 (1): pp. 1–16. doi:10.1109/TKDE.2007.9

[10] C. Fleet, Kowal, K., and Přidal, P. (2012). Georeferencer: Crowdsourced Georeferencing for Map Library Collections. *D-Lib Magazine*, November/December 2012, 18(11/12). doi:10.1045/november2012-fleet http://www.dlib.org/dlib/november12/fleet/11fleet.html

[11] Gemmell, G. (2014). Should Google Be Mapping Tribal Lands? *Daily Beast, TECH + HEALTH*, 4th June 2014. http://www.thedailybeast.com/articles/2014/06/ 04/should-google-be-mapping-tribal-lands.html

[12] Google (2014). *Indigenous Mapping Day is August 9, 2013. Show your support & improve Google Maps!* Google, Map your World Community. https://sites.google.com/site/ mapyourworldcommunity/indigenous-mapping

[13] Halevy, A., Norvig, P. and Pereira, F. (2009). *The Unreasonable Effectiveness of Data*, Intelligent Systems, IEEE , vol.24, no.2, pp.8,12, March-April 2009. doi: 10.1109/MIS.2009.36

[14] Heath, T., and Bizer, C. (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1) 1-136. Morgan & Claypool. http://linkeddatabook.com/editions/1.0/

[15] *In Concert* (2014). accessed 23/6/2014 http://inconcert.datatodata.com

[16] Leverhulme (2014). *Research Project Grants*. The Leverhulme Trust. accessed 23/6/2014. http://leverhulme.ac.uk/funding/RPG/RPG.cfm

[17] McVeigh, S. (1992–2014) *Calendar of London Concerts 1750–1800*. (Dataset) Goldsmiths, University of London. http://research.gold.ac.uk/10342/

[18] Plato. *Phaedrus*, circa 360 B.C.E (tr. Benjamin Jowett) http://www.gutenberg.org/files/1636/1636-h/1636-h.htm

[19] Raimond, Y., Scott, T., Sinclair, P., Miller, L., Betts, S., and McNamara, F. (2010). *Case Study: Use of Semantic Web Technologies on the BBC Web Sites*. Semantic Web Use Cases and Case Studies, W3C. January 2010. http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/

[20] Ramduny-Ellis, D., Dix, A., Rayson, P., Onditi, V., Sommerville, I. and Ransom, J. (2005). *Artefacts as designed, Artefacts as used: resources for uncovering activity dynamics*. Cog. Tech. and Work, 7(2):76-87 http://alandix.com/academic/papers/CTW-artefacts-2005/

[21] Part 2D: Main Panel D criteria, *Panel criteria and working methods*, REF2014, Research Excellence Framework. January 2012. http://www.ref.ac.uk/pubs/2012-01/

[22] schraefel, m.c., Smith, D., Russel, A., Owens, A., Harris, C. and Wilson, M. (2005) The mSpace Classical Music Explorer: Improving Access to Classical Music for Real People. In, *V MUSICNETWORK OPEN WORKSHOP*: Integration of Music in Multimedia Applications, Vienna, Austria, 4–5 Jul 2005. http://eprints.soton.ac.uk/261033/

[23] Sommerville, I. (2010). *Software Engineering*. Harlow, England: Addison-Wesley. ISBN: 978-0-13-703515-1

[24] *API*. Transforming Musicology. Accessed 23 June 2014. http://www.transforming-musicology.org/api/