Crowdsourcing and Scholarly Culture:
understanding expertise in an age of popularism

Alan Dix
Computational Foundry, Swansea University, UK

Rachel Cowgill
School of Music, Humanities and Media, University of Huddersfield, UK

Christina Bashford
School of Music, University of Illinois at Urbana-Champaign, USA

Simon McVeigh
Department of Music, Goldsmiths, University of London, UK

Rupert Ridgewell
British Library, UK

Abstract

The increasing volume of digital material available to the humanities creates clear potential for crowdsourcing. However, tasks in the digital humanities typically do not satisfy the standard requirement for decomposition into microtasks each of which must require little expertise on behalf of the worker and little context of the broader task. Instead, humanities tasks require scholarly knowledge to perform and even where sub-tasks can be extracted, these often involve broader context of the document or corpus from which they are extracted. That is the tasks are macrotasks, resisting simple decomposition. Building on a case study from musicology, the *In Concert* project, we will explore both the barriers to crowdsourcing in the creation of digital corpora and also examples where elements of automatic processing or less-expert work are possible in a broader matrix that also includes expert microtasks and macrotasks. Crucially we will see that the macrotask–microtask distinction is nuanced: it is often possible to create a partial decomposition into less-expert microtasks with residual expert macrotasks, and crucially do this in ways that preserve scholarly values.

1. Introduction

Plato grappled with the way Socrates, his hero and mentor, had been summarily executed by the democracy of Athens; and how easy it is for democracy to slip into ochlocracy and from that to tyranny. In an age when the UK Justice Secretary could publically pronounce that "*people in this country have had enough of experts*" [Go16], how do we in the academe tread the line between expertise and elitism?

In this chapter, we explore the barriers to crowdsourcing within the digital humanities. As digitised sources become ever more extensive, they overwhelm the possibility for complete analysis by traditional scholarship. Crowdsourcing and computational analysis offer ways to deal with otherwise impossible large volumes of material, and yet run the risk of simply creating voluminous trash.

Is academic resistance to crowdsourcing an elitist fear of the unwashed, or justifiable wariness of incipient poor scholarship?

We will attempt to dig into some of the core values that lie at the heart of scholarly culture, exploring how issues of authority and integrity are crucial not to the maintenance of the scholarly elite, but to the nature of scholarship itself. Through this understanding we explore ways in which digital technology could allow wider participation whilst preserving the core values of academia.

As a case study, we draw on our experience in a particular domain: the study of the development of public musical performances through evidence of ephemera, such as notices and advertisements, and our work to create a definitive digital archive in the *In Concert* project and earlier projects.

As an academic domain, this stands in contrast to more traditional musicological approaches that place composers, performers, patrons and critics – the elite of the music world – at centre stage. Instead, the focus on audiences, performance, ephemera and the development of print-music consumption is one that gives voice to the listener, and, to an extent, the masses.

However, taking seriously the role of mass print-culture as the subject of study does not mean these studies themselves are not expert activities. Indeed, the plethora of long-dead performers and now-a-days obscure composers makes the area opaque to all but the most knowledgeable. When creating a scholarly digital archive, throwing open anything but the most mundane activities to crowdsourcing appears to risk polluting the authoritative corpus.

Within the bounds of the *In Concert* project we have not fully managed to square this circle, but we have been able to combine varying levels of expertise and automated contributions as part of a reimagined process of digital archive creation. Through this we believe we have come closer to understanding potential ways forward, including critically the use of digital infrastructure to maintain adequate provenance to ensure that when data is viewed its authoritative, or non-authoritative status is evident. This parallels lessons from scientific crowdsourcing, which use a variety of means to develop measures of expertise, trust and degrees of certainty.

In the remainder of this chapter we will first look at digital archives, and crucially the way the dichotomy between macrotasks and microtasks is less clear when we consider the way expert macrotasks can be decomposed for crowdsourcing. We then proceed to describe the key case study for the chapter, the *In Concert* project, including its datasets, and some of the barriers to progress it has encountered. We consider the potential to address some of these barriers using crowdsourcing or automation, both in general within the digital humanities and considering sub-tasks within *In Concert* itself; however, we will see that crowdsourcing brings its own problems and barriers. Some of these barriers to crowdsourcing are technical, but some more fundamental, about the nature of the academic process, and so we then look at the scholarly values and academic value mechanisms that drive and constrain work in the humanities. By understanding these, the *In Concert* project was able to effectively employ automatic and non-expert human processes in various substantive sub-tasks. By studying these successful applications of non-expert, but not crowdsourced, interventions, we develop heuristics that have the potential to encourage and enable appropriate macrotask crowdsourcing in the humanities.

2. Crowdsourcing of Digital Archives

Crowdsourcing has already been effectively used in the digital humanities, for example in projects inviting members of the public to align historic maps with current maps.

However, it is also clear that some aspects of digital humanities are not easily amenable to crowdsourcing. Interpreting a 13th century letter may require not only an understanding of the language and writing style of the time, but also an appreciation of the political and personal relationships within court. This is evident in even relatively short time scales, for example, the mutation of the word 'celebrity' from the quality of a solemn occasion to a B-list reality TV star.

The case study in the chapter concerns the creation of digital archives, many dating back to just the 19th century, so with fewer linguistic barriers than older material, but still requiring scholarly expertise and knowledge of the time, personae and available repertoire.

2.1 Corpus creation process

Figure 1 shows a simplified view of the process for the creation of digital archives.

Stage 1 is the low level digitisation/transcription and clearly most amenable to either automation of crowdsourcing via microtasks as, for relatively modern sources, they require little expertise beyond normal language skillset.

Stage 2 includes more complex tasks, which require more expertise. It is at this stage that the academic value of the digital corpus is largely created. The tasks even here range from those requiring deep knowledge of the period or subject matter, and some that, at first sight, may involve less expert knowledge.

The output of this second stage is an authoritative digital archive that can be used as a base resource for further scholarship leading (stage 3) to publication: books, chapters and articles. Typically this may first be carried out by the scholars who produced the article, but then later the authoritative archive may be released to those outside the boundaries of the original team or institution.

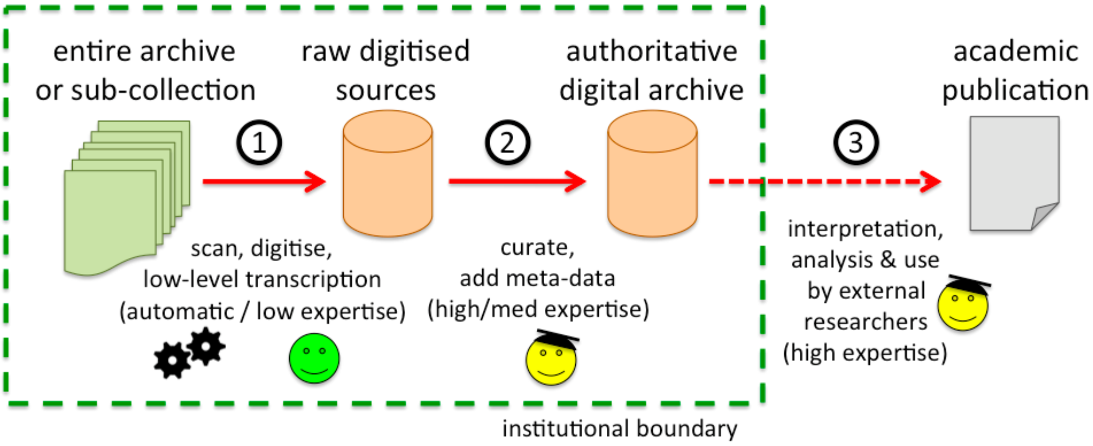

entire archive or sub-collection
raw digitised sources
authoritative digital archive
academic publication
1
2
3
scan, digitise, low-level transcription (automatic / low expertise)
curate, add meta-data (high/med expertise)
interpretation, analysis & use by external researchers (high expertise)
institutional boundary

Fig. 1. The digital archive process (from [DC14])

In this chapter we will be focusing most extensively on stages (1) and (2) and perhaps most crucially stage (2), which emerges as a bottleneck in the *In Concert* case study.

2.2 Macrotasks and microtasks in corpus creation

Figure 2 shows different kinds of task along the axes: the *size* of individual items of the task; and the *expertise* needed to accomplish the task.

At the top left (A), we have large tasks requiring little expertise, for example, given a 1950s map and 1970 map of London align the locations of road junctions common to both. At lower left (B) we have small inexpert tasks, for example, extracting the item and cost from a single line a receipt. At upper right (C) we have large expert tasks, for example, understanding the correspondence of a minor poet. Finally at the lower right (D), we have small tasks requiring expertise, for example, in a single paragraph of a correspondence identifying the names of other poets of the time.

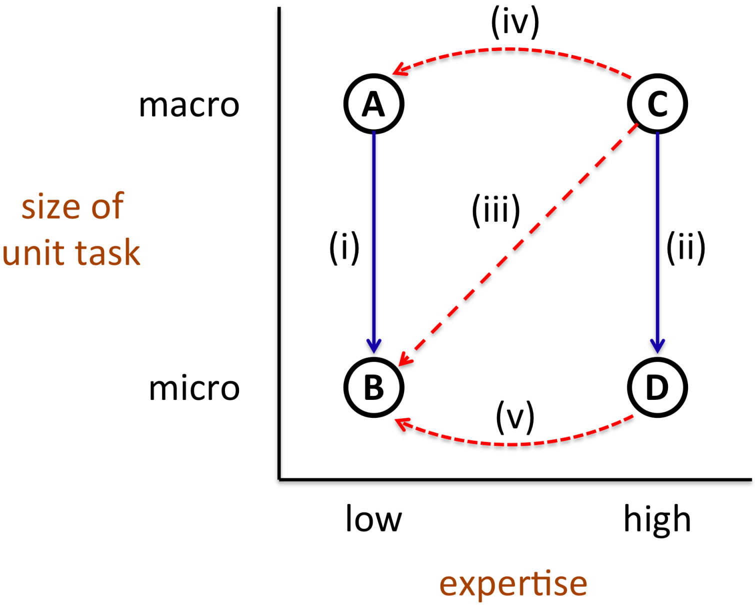


Fig. 2. Expertise and task decomposition

Traditional crowdsourcing is effectively about the move (i) from (A) to (B), breaking down a large tasks into small parts, each able to be assigned individually to relatively inexpert workers. In contrast, traditional professional work often involves a level of task decomposition (ii) from (C) to (D); indeed this is precisely the purview of classic time-management techniques.

Of course, this is a simplification. There are many gradations of expertise, and we will see examples where there is a distinction between work that can be carried out by junior academics, and work that requires a field expert.

In the digital humanities we will typically start with large expert tasks (C), and ideally would like to break it down into many small microtasks that are amenable to low-expertise crowdsourcing (A). That is we would like to make transition (iii).

In the simplest case once the decomposed microtasks are performed, the overall macrotask is itself complete; for example, if we have transcribed each individual phrase of a speech, we have transcribed the whole speech. However, at very least there is a level of automatic processing, to aggregate the results of the microtasks. Furthermore, there are often residual macrotasks that need to be performed (fig 3); for example, in the map-matching task there may be discrepancies due to crowdsource worker errors, complexity of the data (e.g. two streets or landmarks with the same name), or errors by the original map-maker. Often these residual macrotasks involve greater expertise than the crowdsourced microtasks, but are easier or less voluminous once the microtasks are complete . Effectively this is achieving transition (iii) by route (iv)–(i).

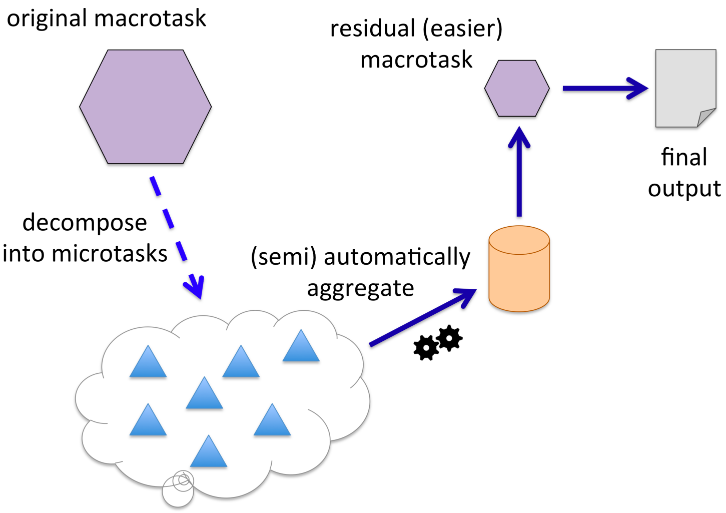

original macrotask
residual (easier) macrotask
final output
decompose into microtasks
(semi) automatically aggregate

Fig. 3. Residual expert macrotasks.

In fact the three stages of figure 1 can be seen as an example of transition (iii) by route (iv)–(i). The highly expert tasks of creating scholarly outputs is broken down into three stages, the first of which requires less expertise than the latter two. Furthermore, stage 1 is often amenable to decomposition into microtasks even if these are at the junior academic level rather than full crowdsourcing.

The other potential route from (C) or (B) is via route (ii)–(v). The initial expert macrotask is first decomposed into many expert microtasks and then each microtask is further decomposed into a less-expert and more-expert part (Fig 4). The less-expert part may then be amenable to crowdsourcing, automatic processing or delegating to junior academics. Many of the examples we shall encounter in the *In Concert* case study fall into this pattern

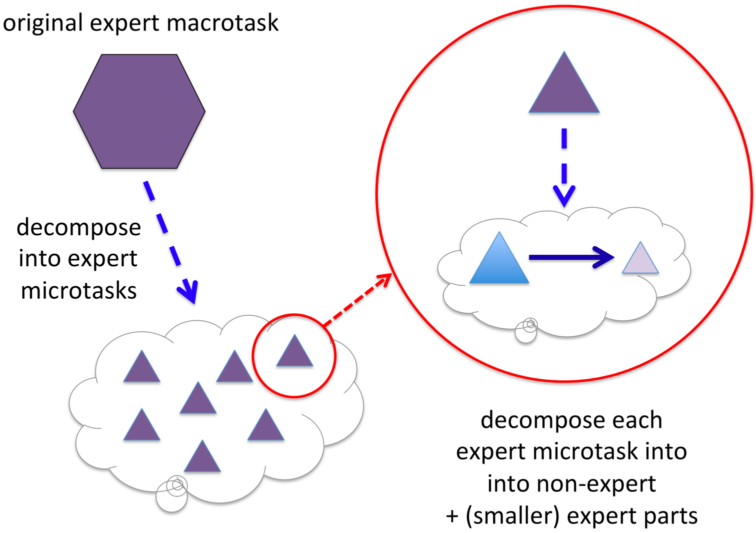

original expert macrotask
decompose into expert microtasks
decompose each expert microtask into into non-expert + (smaller) expert parts

Fig. 4. Decomposing microtasks.

Often this microtask decomposition may be in the form of the expert microtasks that simply verify the initial less-expert microtask. There may also be some form of pre-filtering into simpler and harder macrotasks, or some form of validation that highlights discrepancies, or other cases requiring more expert interventions.

2.3 The myth of the acontextual

Haas et al. [HA15] describe a microtask in terms of questions that "require little context or training to answer". We have discussed the 'little training' aspect in terms of expertise, but the acontextual element also requires examination.

Clearly some tasks require an understanding of a whole corpus, for example, assessing the mood of a particular politician from reading correspondence written during the lead up to a critical event. There are crowdsourcing techniques targeted at such non-decomposable tasks. Notably TAS (Task Assignment and Sequencing) passes large tasks of this kind sequentially between a number of crowd workers; each spends considerable time on the task as a whole, advancing work on it, before passing on to another [SL18]. An R&D tasks was used for the empirical evaluation of TAS, which also included a pre-test for knowledge of the domain (e.g. FIFA); that is an element of crowdworker expertise.

In many tasks, the decomposable/non-decomposable distinction is less dichotomous than first appears. Think of the map-matching tasks. Maps may vary in the way they portray different features, for example, showing built up areas as blocks of colour or divided into individual properties; or they may use different abbreviations. Although each atomic matching task is relatively independent, still there will be a level of learning as the task is performed.

Sometimes, the microtasks only make sense within the larger context, for example, we will see in fig. 9, how one of the musicologists in In Concert spreads out paper across a table as part of what appears to be a more focused matching task.

Furthermore, some of this learning is likely to feed into higher-level understanding. Spending time identifying key names and events from a politicians letters may seem like a low-level task, but of course is immersing the reader in the life of the writer. Indeed, one method for dealing with creative tasks is to deliberately create a 'busy work' aspect, which can be performed fairly automatically, but is at the same time orienting one's mind towards the larger creative task [Dx19].

Any outsourcing of microtasks to crowdwork or automation needs to be cognisant of these subtle, but crucial effects (Fig. 5). For example, very early CAD systems were introduced in architects offices in the late 1970s in order to reduce the time consuming tracing of plans from previous projects, which was often the first stage in starting a new related project. Although it certainly sped up the drafting process, the architects found themselves more highly stressed and less productive overall: the low-level tracing activity had been giving them precious time to think about and prepare for the creative task.

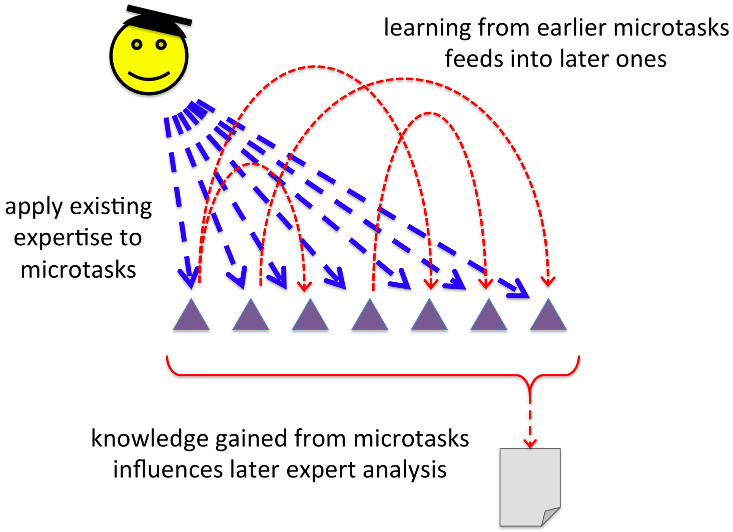

learning from earlier microtasks
feeds into later ones
apply existing
expertise to
microtasks
knowledge gained from microtasks
influences later expert analysis

Fig. 5. Microtasks lead to understanding.

3. The *In Concert* Project

This chapter draws on case studies from *In Concert: Towards a Collaborative Digital Archive of Musical Ephemera* [IC16], a sub-project of the AHRC funded *Transforming Musicology* programme [TM16]. This project was a collaboration between Musicology and Computer Science and had a dual aim. On the one hand the musicology goal was to enhance a number of datasets related to concerts in London from the 18th century onwards. However, there was also a broader digital humanities goal to use this experience to better understand the evolving role of the digital archive. Indeed, this chapter is one of the outcomes of this broader goal.

3.1 Performance and ephemera

Much of musicology is focused on composers and their works. This may include historical study of the lives of the composers and of the development of individual works from sources such as letters, contemporary accounts and official records. In the way of academia, this involves highly specialised study of relatively sparse sources.

In contrast *In Concert* was focused on actual performance of music – what was selected and listened to rather than what was produced. The canon, the works that persist, is not merely about the 'genius' of the individual composer, but also the trends within broader culture. Furthermore, the patterns and trends of performance and performer are not merely reflections of taste, but also connect to issues in social history such as the role of women and the privileging of repertoire that reflects the interests and identity of the culturally empowered [CR12].

Today, the consumption of music may be studied directly and near instantaneously through streaming services such as Spotify, downloads from iTunes, the schedules of BBC Radio, or even, for popular music, the long running weekly 'top 20'. However, in pre-internet days, the sources are more diverse and dispersed, often in the form of *ephemera*: concert programmes, newspaper reports, and advertisements; things never intended to be preserved for posterity (although historically concert programmes were often collected and bound).

As with any historical source, there is partiality and bias in what was reported and what was preserved: the concerts of high society are more visible than the songs sung in taverns. However, to give the most reliable picture of the patterns of performance, the ephemera needs to be sampled, collated and recorded in a consistent and scholarly manner.

3.2 Datasets

In Concert focused on three principal datasets:

- LC18 – *Calendar of London Concerts 1750–1800* [MV92] – This was created from many sources relating to concerts in the second half of the 18th Century. Given the relatively small number of sources and events during the period it is a near exhaustive collection of available information.
- LC19 – *Concert Life in Nineteenth-Century London* [BC00] – By the 19th century the number of concerts and relevant print sources grew to such an extent that a complete compilation is not possible; instead, sample years at 20-year intervals were exhaustively studied, using newspaper archives and other sources.
- CPE – *Concert Programme Exchange* (*Konzertprogramm Austausch*) 1901–1914 – In the early years of the 20th century Leipzig publisher Breitkopf & Härtel distributed printed copies of programmes of major concert venues in Europe, Russia and America. The British Library's collection of these was digitised for Gale Cengage, making around 12,000 programmes available in OCR form.

These were supplemented with two other datasets primarily as authority files:

- CPP – *Concert Programmes Project* [CP04] – This project, administered at the British Library, collates meta-information about archives; it does not contain programmes or programme text itself, but lists archives and collections (most offline) where such ephemera can be found and information about the venues and people they cover.
- BMB – *British Musical Biography 1897* [BS97] –This 400 page volume includes nearly 4000 entries for British musicians and composers during the nineteenth century and is broadly contemporary with the first editions of Grove [Gr00]. A digital version was created as part of the *In Concert* project based on OCR from the Internet Archive.

These data sources overlap in terms of subject, people and venues, but represent very different stages of digitisation from raw OCR (CPE, BMB) to fully authoritative corpus (LC18) and from full details of individual concert programmes (CPE) to meta-information about presence of archives (CPP).

3.3 Barriers to progress

One of the drivers for the *In Concert* project was a hiatus in the development of the LC19 dataset. As noted the sources for the 19th century are far more extensive than used for LC18. The 1750–1800 dataset LC18 had been the work of an individual, whereas LC19 required a team project including three primary investigators and a substantial number of research assistants.

Funded projects [CL97] in the mid-late 1990s were used for a first phase of the LC19 development. A relational database structure was created based on the experience of LC18 and this was initially populated by the research assistants extracting information from primary sources, principally newspaper adverts and notices and concert programmes. The research assistants would create a record for each advert/notice and fill in details such as the date, venue, performers, works and composers in the programme. This was successfully completed and this first phase data was used as the basis of initial analysis and publications [BC00].

However, a second phase was always envisaged. The initial data collection was effectively 'raw' data: one entry per notice, and raw text names of people and venues. The plan was to create a dataset with a single entry per concert, critically editing partial information from multiple notices, linking people, venues, works etc. to unique authority identifiers (e.g. if there were several variants spelling of the same person's name, or several distinct people shared a common name).

This second phase would have allowed better connections with the LC18 dataset and also statistical analysis of historical trends, visualisations, etc. In particular LC19 has authority files (people, venues, works), which could be connected to the authority files in LC18.

However, this second 'interpretative' phase required more expertise and professional judgement, and so needed the time of the more senior academics, which of course is limited. Consequentially progress on this second phase had stalled for some time.

4. Opportunities for Crowdsourcing and Automation

4.1 Challenges of scale

The difference between the LC18 dataset and LC19 demonstrate the challenges of scale inherent in digital humanities research. LC18, with about 4000 concerts, was already an extensive exercise, but was possible by a single dedicated scholar. However, the increase in the number of popular performances and available sources in the 19th century meant that even a one-year-in-twenty sample required a substantial team effort and its final phase was incomplete. Even this belies the fact that the volume was changing throughout the 19th century, so that, by the time we come to the 20th century, archives of individual concert houses are themselves of similar or greater extent and CPP's meta-records of these archives are themselves large.

This increase in volume is a result partly of a greater number of events, but also the greater number of preserved sources, the problem of the 'infinite archive' [Be04]. For classicists or traditional scholars dealing with sparse sources, the problem is interpreting the little data that is available. For 'born digital' materials, such as Spotify logs, the issues are almost those highlighted by Borges' imaginary map [Bo46] where the data is almost coincident with the world itself; given the massive volume data the problem is what to select or even ignore in order to turn raw data into information.

Between these extremes are areas such as those dealt with in *In Concert*, where the number of physical or raw digitised resources is too great for scholars ever to deal with, and yet requires a level of processing and interpretation before it is suitable for scholarly analysis.

Note this does not invalidate traditional scholarly approaches to historical archives. If you have a focused topic of study such as the works of a minor composer, or performances in a particular venue, you still need to trawl through multiple archives to find heterogeneous sources. Although even raw digitisation may make this easier allowing faster searching and less travel to view originals if not necessary, and certainly avoiding speculative journeys only to find there are no useful resources.

However, it is a problem for the scholar wishing to study broader questions such as different patterns of repertoire between European centres, or the changes in musical taste in London venues during the 19th century.

Sampling, as in the LC19 dataset, is a partial way to deal with this issue, but, as we have seen, even a high level of sampling can still lead to datasets too large for expert scholarly curation.

This impasse appears to create an obvious opportunity for crowdsourcing or automated solutions.

4.2 Candidate tasks

Looking through the various datasets we can identify a variety of tasks requiring differing levels of expertise and hence potentially offering opportunities.

T1 – Low-level digitisation. – This may require special equipment for high-quality photography or scanning, but also may include tasks such as transcription, or correction of OCR. For example, in CPE the title pages of concert programmes often use decorative scripts, which are hard to OCR.

T2 – Identification of format or general language features. – For example, in BMB the transcription included page headers, capitalised entry names, etc. In CPE concert programmes often included columns of names.

T3 – Identification, marking or extraction of semantic fields. – For example, in LC19 finding the name of the venue in a newspaper advert. Another example in LC19 was the initial identification that a portion of a newspaper page was in fact a concert notice.

T4 – Matching text names of venues, people and works to unique entries in authority files. – For example, there may be two John Smiths, father and son, who can be disambiguated by the date of the concert or the style of music. This task might also require knowing that certain performers had multiple stage names, or that a venue changed name.

T5 – Matching authority files between datasets. – The LC18, LC19, BMB and CPP all include unique identifiers of performers and composers and LC18, LC19 and CPP include venue identifiers. By connecting these not only is it possible to analyse the datasets together, but also where one dataset includes information such as external identifiers or geocoding, this becomes shared by the other datasets reducing work.

T6 – Grouping notices (in LC19) that refer to the same concert. – In some cases this may simply be that two notices refer to the same venue on the same date, but some venues are large enough to have several concerts on a single day, also some notices may be vague about times, may have errors, or dates may change if a concert is postponed. In short even the most simple concert notice/programme often has rich many-to-one relational complexity.

T7 – Merging groups of notices into a single definitive concert record. – In some cases this may simply be filling in details that are missed in one notice with complementary information in another. However, on other occasions this may require choices between conflicting information.

T8 – Musicological analysis of the dataset. – This may be by hand or by using data processing, statistical, or visualisation techniques.

Looking back to figure 1, tasks T1–T3 belong roughly to stage 1, T4–T7 to stage 2 and T8 corresponds to stage 3. It is clear that some of these tasks require less musicological expertise than others. In LC19's first phase the research assistants performed T1 and T3 (and T2 where relevant) but T4, T6 and T7 were left for more expert processing in phase 2.

4.3 Barriers to crowdsourcing – low-level

Transcription or correction of OCR sound like straightforward candidate microtasks for crowdsourcing. However, it is interesting that the raw OCR text of BMB, (and similar documents) at the Internet Archive appears uncorrected. This appears to be partly related to complexity. It is possible for readers to correct OCR and then upload corrected versions, but this really requires a volunteer to commit to correcting all, or a substantial part of a volume. This complexity barrier has been partly addressed by other projects.

reCAPTCHA was originally used on the New York Times archive, and it was proposed in some reports that it could be used for the Internet Archive [vA08], but it is not clear whether this ever occurred before reCAPTCHA was acquired by Google.

Distributed Proofreaders [DP18] is a web-based service set up originally to help volunteer correction of Project Guttenberg texts. It allows page-by-page correction and manages different stages of proof correction from first OCR scans to more complex verification. However scanning the title of volumes processed, it is evident that the majority are either novels or books of a largely textual nature (e.g. [Ac22]). Tomes such as the British Musical Biography or gazetteers are less obvious candidates for the volunteer.

Historical texts are also harder to OCR (less distinct fonts, poorer quality paper and printing, non-standard spellings). In the case of concert programmes and notices, a great deal of information is also communicated via changes of font and tabular positioning on the page, similarly catalogue-style books such as directories, dictionaries and gazetteers often include abbreviations and special conventions, some of which, such as bolding, may be difficult to retain in OCR. There have been projects to create special-purpose OCR tools and workflows for historical texts, for example the PoCoTo open source software [VG14] and Fink et al.'s system to create adaptive OCR based on previous proof corrections [FS17]. However, to date these are not part of the Internet Archive's standard workflow.

4.4 Barriers to crowdsourcing – more complex tasks

As noted, the hiatus in the LC19 dataset was at a stage way beyond these low-level tasks. Academically trained research assistants read physical or digital copies of newspapers, found references to concerts and then extracted all available relevant information to input into the SQL database. This was already deemed a task requiring a level of academic expertise and training to use the database, although some aspects of the tasks might well have been possible to crowdsource (e.g. locating concert notices).

However, even the research assistants were not deemed sufficiently expert to perform tasks T4, T6 and T7 on the LC19 dataset. To an outsider aspects of these tasks look as though they could be suitable for crowdsourcing. For example, T6, grouping multiple notices that relate to the same concert, appears to be something that is possible based on general knowledge and understanding: looking through date ordered lists of notices, and collecting those that appear to be at the same venue at the same time.

Early in *In Concert* the potential for using knowledgeable amateurs for crowdsourcing was discussed. These were often referred to as 'Radio 3 listeners' – Radio 3 is the BBC classical music radio channel in the UK, and listeners tend to come from both a slightly older and more highly educated demographic than the general population. The general idea of using such knowledgeable crowdsourcing was accepted as a good idea, but any suggestion of actually doing this for specific tasks in the dataset was greeted with concern, the idea of the musical amateur seeming to be at odds with that of the scholarly corpus'.

Again, looking from the outside, this at first may seem to be a case of excessive scholarly purity. However, digging deeper it relates to justifiable caution – any uses of crowdsourcing for macrotasks needs to be done in ways that understand and fit within the overall scholarly culture. We should note that we were not the first music-based project to have to deal with problems in this area [BL12].

5. Scholarly values and academic value

Key to the success of any system deployment whether digital, physical or organisational, is an understanding of the underlying values and value within the setting.

- *Individual values* – What are the internal beliefs, motivations and drivers that create a sense of personal worth and lead individuals or groups to judge the worthwhile nature of outcomes?
- *Value mechanisms* – What are the external measures, rewards and validation offered by the wider system in which individuals or groups participate?

Those entering academia on the whole assent to a number of common scholarly values such as integrity and the desire to increase the bounds of scholarship. However, they also operate within a complex of reward and career advancement mechanisms including promotion procedures, metrics for external assessment (such as the UK REF), and publication routes.

In previous work we have explored the values and value mechanisms that are critical in forming attitudes towards crowdsourcing and automation within digital humanities [DC14]. We will summarise these as a basis for understanding potential ways forward.

5.1 Scholarly values: authoritative and complete

The term *authoritative* in the above is crucial both for the scholar's own use and for the scholar to be happy for others to see the work. The methods of creation need to be well-documented and of consistent high quality so that further scholarship can be built upon it.

In some cases the corpus may not be exhaustive, but it is important that it is complete in the sense of covering a known period, geographic area or other selected (and stated) criteria. This may include sampling, as has been done with LC19, but in this case the sample needs to be unbiased and clear in its criteria.

In essence this is about the ability of the scholar using the corpus to be able to assess the reliability of data within it and make defensible inferences and arguments based upon it.

Any dataset inevitably embodies potential bias in the collection and preservation methods (as noted history selects for the rich and powerful) and also in interpretation. Indeed scholars differ in their approaches to the record and each scholar's use of a resource will vary depending on their assessment of the curator's hermeneutic.

In this context, a distrust of the amateur is understandable. If a known scholar has curated a digital archive, then those using it can take this into account; even if they disagree with the curator, they can still rely on basic levels of scholarly consistency and accuracy. If many amateur hands are at work during crowdsourcing it seems impossible to know if all of the data is of sufficient quality without checking everything, and furthermore different workers may make inconsistent decisions.

As well as potential problems in the use of the resulting corpus, those in charge of curating the digital archive feel responsible for it. If there are inaccuracies or omissions, they will feel they are letting down their own personal standards and potentially weakening their academic credibility and reputation.

5.2 Academic reward

Stage 3 of Figure 1 includes the digital archive being available to the wider research body as well as the scholars involved in its creation. In practice this may be delayed for many years, or even indefinitely.

One reason is related to the scholarly values above: the curator(s) need to be very sure they are releasing a corpus on which they feel comfortable to rest their scholarly reputation. Preparing a corpus to the point where you can perform your own research is less onerous as you understand the limitations and sources of various parts, and so are able to make assessments of validity.

Intellectual property issues are also problematic: some sources restrict access to personal research, one's rights to republish derived datasets may be unclear, and it may be hard to determine the correct licence under which to release one's own data.

Technical barriers may also deter publication of data. Although this is becoming easier as many universities create digital repositories, the complexity and costs of digital archiving are perhaps underlined by the UK Arts and Humanities Research Council (AHRC). During the 2000s the AHRC mandated that all funded projects lodge their resulting data in the AHRC's own archive, the Arts and Humanities Data Service, possibly in the process leading researchers to believe this was an archival store backed by the resources of government. However, by the end of the decade the AHRC not only dropped the requirement, but closed the repository [Ri07, Wi19].

This story underlines the ambiguous role of data in the research process and hence the most critical reason for delaying dataset publication.

Broad scholarly values lead one towards openness, expanding the breadth of knowledge. However academic reward mechanisms both formal and informal are oriented primarily towards scholarly publication in books or journal articles (depending on the discipline). Although the community will be grateful to the scholar who makes curated resources available, the real academic applaud goes to the scholars who interpret those resources and create publications from them.

In the UK Research Excellence Framework, the periodic assessment of national academic research, Panel D, which covers arts and humanities, did include a curated 'database' as a valid research output [REF12]. However Panel B (science and engineering) did not mention data as a valid output at all, despite the Web being developed by Berners Lee precisely to share scientific data from CERN [BL89]. The Leverhulme Trust, which funds cross-disciplinary research is even more specific explicitly rejecting applications where "*the balance between assembling a data bank or database and the related subsequent research is heavily inclined to the former*" [LT18]

In summary, academia regards the publication of data as *valuable*, but does not *value* it.

6. Radical transformations to support traditional values

Having disentangled some of the complex web of values and reward mechanisms that underlay the scholarly curation process, our challenge within *In Concert* was to radically re-imagine that process in ways that preserve the underlying scholarly values and work within the academic reward mechanisms and yet are more open in terms of both publication of data and accepting automated or non-expert input.

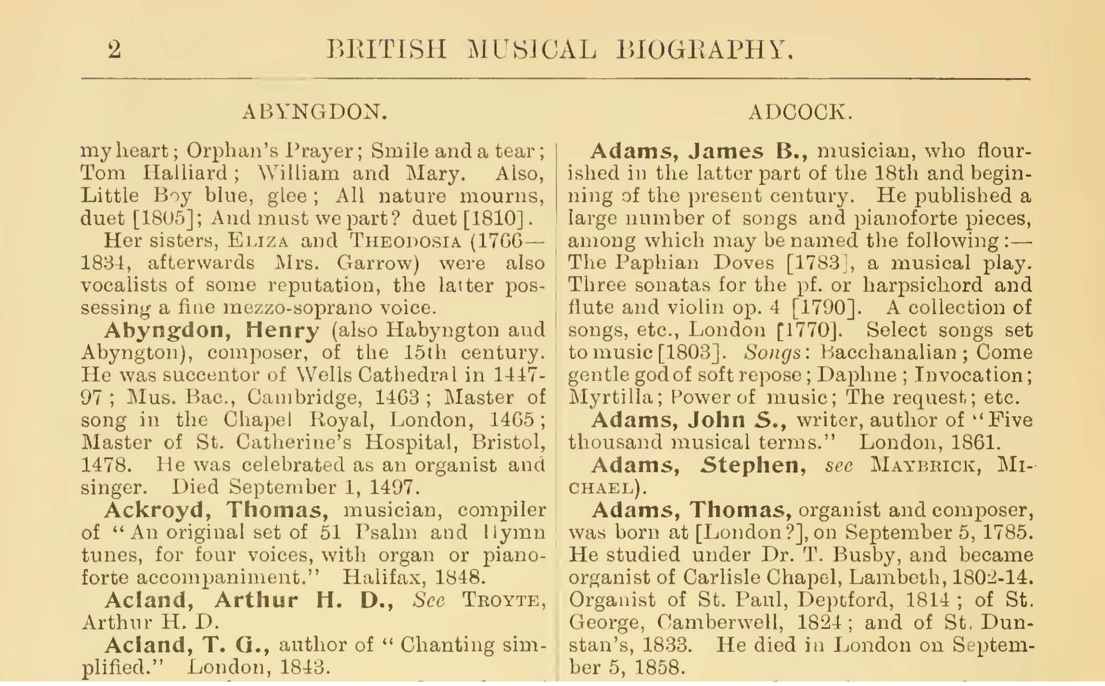
As noted earlier in this chapter, we did not adopt crowdsourcing. This was partly for reasons of time, and partly because the team was still resolving the issues and barriers as described. However, we did use automated algorithms, which on the surface have some similar problems to crowdsourced work, and also two human non-experts (a very small crowd), the technical partner in the project and another non-expert known to the team.

We will describe three tasks in the project where these non-experts (human and machine) formed part of the process and then return to reflect on the lessons this has for future crowdsourcing of macrotasks in the humanities. Each of these follow broadly the route (ii)–(v) outlined in section 2.2. Each takes an initial macrotask, creates a combination of non-expert microtasks, semi-independent expert microtasks, and residual expert macrotasks.

6.1 A digital version of the British Musical Biography

One of the core datasets of *In Concert* was CPE, the *Concert Programme Exchange*, which was at the earliest stage of preparation with OCR only. The range and complexity of the documents, concert programmes from many venues, meant that further automatic processing would be very difficult. It is an ideal candidate for both low-level crowdsourcing, tidying up OCR of florid fonts, and also higher-level tasks such as marking up titles, players of different instruments, pieces performed, etc.

The British Musical Biography (BMB) was at a similar stage of preparation, with a raw OCR at the Internet Archive, but its strong structure made it far more amenable to automated processing (see fig. 6). This was valuable in its own right, but, more important, acted as an exemplar allowing the project to learn lessons and develop processes which, we hope, would be useful for the more complex CPE.



2 BRITISH MUSICAL BIOGRAPHY.

ABYNGDON.

my heart; Orphan's Prayer; Smile and a tear; Tom Halliard; William and Mary. Also, Little Boy blue, glee; All nature mourns, duet [1805]; And must we part? duet [1810].

Her sisters, ELIZA and THEODOSIA (1766—1834, afterwards Mrs. Garrow) were also vocalists of some reputation, the latter possessing a fine mezzo-soprano voice.

Abyngdon, Henry (also Habyngton aud Abyngton), composer, of the 15th century. He was succentor of Wells Cathedral in 1447-97; Mus. Bac., Cambridge, 1463; Master of song in the Chapel Royal, London, 1465; Master of St. Catherine's Hospital, Bristol, 1478. He was celebrated as an organist and singer. Died September 1, 1497.

Ackroyd, Thomas, musician, compiler of "An original set of 51 Psalm and Hymn tunes, for four voices, with organ or pianoforte accompaniment." Halifax, 1848.

Acland, Arthur H. D., *See* TROYTE, Arthur H. D.

Acland, T. G., author of "Chanting simplified." London, 1843.

ADCOCK.

Adams, James B., musician, who flourished in the latter part of the 18th and beginning of the present century. He published a large number of songs and pianoforte pieces, among which may be named the following:—The Paphian Doves [1783], a musical play. Three sonatas for the pf. or harpsichord and flute and violin op. 4 [1790]. A collection of songs, etc., London [1770]. Select songs set to music [1803]. *Songs*: Bacchanalian; Come gentle god of soft repose; Daphne; Invocation; Myrtilla; Power of music; The request; etc.

Adams, John S., writer, author of "Five thousand musical terms." London, 1861.

Adams, Stephen, *see* MAYBRICK, MICHAEL).

Adams, Thomas, organist and composer, was born at [London?], on September 5, 1785. He studied under Dr. T. Busby, and became organist of Carlisle Chapel, Lambeth, 1802-14. Organist of St. Paul, Deptford, 1814; of St. George, Camberwell, 1824; and of St. Dunstan's, 1833. He died in London on September 5, 1858.

Fig. 6. Portion of Brown and Stratton's British Musical Biography [BS97]

Page beaks were not marked in the OCR, but the page number and capitalised 'BRITISH MUSICAL BIOGRAPHY' were a (relatively) easy marker for automated pagination, similarly the capitalised column headings made them (relatively) easy to spot automatically. The bold font was not marked in the OCR, but the entries are of the form:

Name, Name {optional initials)

where the names have initial capitals and there are a small number of variations. This allowed the entries to also be identified.

These automatic structuring rules were supplemented with sanity checking rules, for example, verifying that page numbers are consecutive and entry names in alphabetical order.

If the original text and OCR had been perfect, this would have enabled computational algorithms to process the text unaided. However, this was not the case. The quality of the print led to frequent OCR errors, for example some capitals (such as 'C') could be read as lower case, lower case L as a bar '|', and commas and full stops could be confused. Added to this there were some errors in the text itself such as comma/full-stop mistakes in typesetting and names out of proper alphabetic order. Finally, although most names fell into simple patterns, others, for example royalty, required specialised rules.

Where failures in sanity checks were attributable to incorrect OCR, the OCR text was edited by hand and the files re-processed. Other failures led to refinements of the rules, for example different name formats. However, in some cases *exception files* were created, that is tables of specific rules such as: "the entry on line 27 of the right hand column on page 23 should read Doe, John". These exception files have become a recurrent pattern in our attempts to automate different forms of processing: not everything can be captured in generic rules.

Finally, a page-by-page check was made to verify that the database entries did correspond to those in the OCR, although there was no attempt to completely fix the OCR in the text within an entry.

It should be noted that the hand checking was carried out completely by the non-experts, and would almost certainly have been possible as a crowd-sourced exercise.

There are a number of factors that made this a possible task for non-experts:

1. The authoritative nature of the work was actually carried out by Brown and Stratton in the 19th century, this exercise was merely a digitising of an existing scholarly resource. Although the kinds of checks and rule creation varied in complexity, there were therefore no scholarly judgements required.

2. Furthermore, because this was not the musicologists' own scholarly work, and merely a digitisation exercise, there was little risk of the work reflecting badly on the scholars reputations.

3. The non experts were known by the team and trusted to be meticulous, for example not correcting apparent misspellings in the text as printed, merely ensuring that the digitised form corresponded to the page.

6.2 Cross-linking authority files

We had name and place information from four sources. Both the 1750–1800 and 19th Century London Concert datasets (LC18 & LC19) have authority files for people (composers and performers) and places (venues). The Concert Programmes Project (CPP) has large authority files for places and agents (people, groups and organisations), including some geo-referencing and planned VIAF links. British Musical Biography (BMB) has people's names only, but is comprehensive.

Automatic matching was used to create candidate matches followed by a hand verification stage. The latter was crucial as the authoritative nature of the data was a key academic value for the humanities researchers [DC14]; automatic matching, whilst useful, is bound to be inaccurate, yielding both false positives and false negatives. Following the principles of 'appropriate intelligence' [DB00], the automatic algorithms were not designed to be as clever as possible, but instead to be part of a human–computer system that as a whole yields reliable results.

6.2.1 Automatic matching

Places were simplest to match automatically using plain word matching and permuted word indexes for efficiency. There are fewer place names than people's names and they tended to be more standardised; so simple matching was sufficient for candidate identification

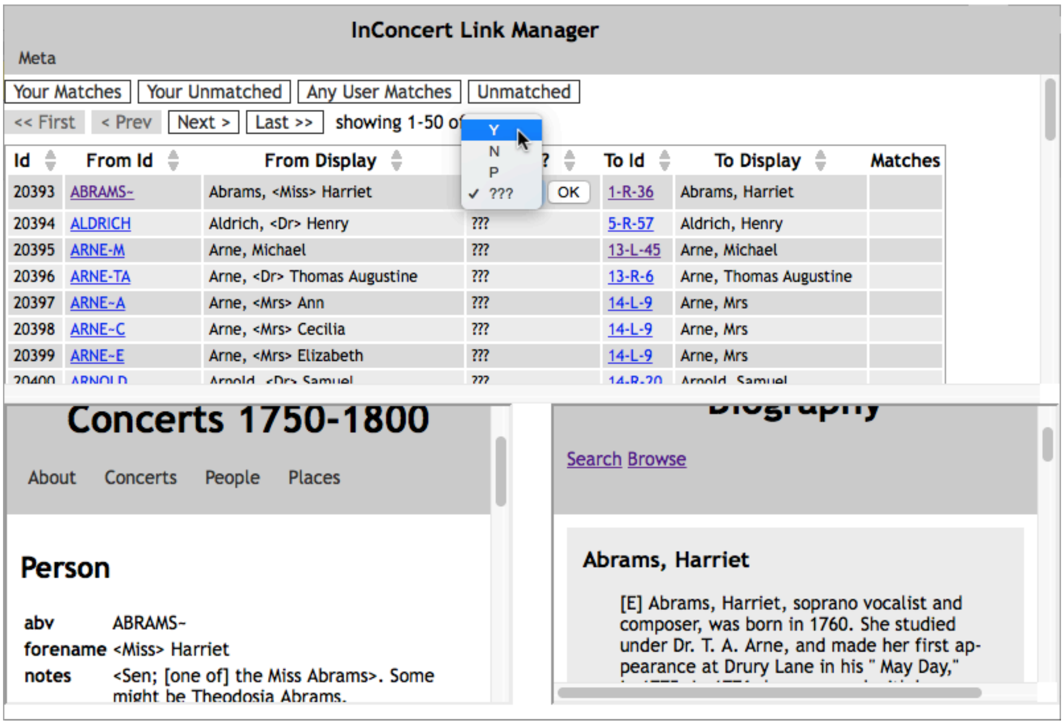
People names were more complex. First, this was because the data sources needed an element of cleaning/normalisation. In the LC18 dataset, the ids included an encoding of the surname, gender and possible disambiguation; for example

"KNEISEL~" for the female (trailing tilde) "Henriette Kneisel", or "TURNER-2" for one of two "Turner"s. This was relatively straightforward pattern matching. More complex was the CPP data, which included groups and organisations as well as people and also was itself garnered from multiple sources. Some people's names had the forename as a separate field, some were in 'surname, first name' format, and some were more complex, including honorifics. In the spirit of maintaining the original source as 'golden copy', this task was managed through a combination of keywords for terms in organisations (e.g. 'orchestra', 'Staatstheater'), extensive lists of honorifics (e.g. 'Prince', 'Mlle', 'Duke of'), and explicit exceptions (e.g. that record id '2173' named 'Tate Britain' is an organisation not someone with surname 'Britain').

Having normalised names as much as possible, the automatic algorithm matched between datasets using a similar word match measure to the places. Fuzzy matches were not used, as this led to too many false positives and the point of the algorithm was to aid not replace human matching. Note that while crude whole word matching was used for the batch processing for names, fast fuzzy search is enabled in online datasets using both Soundex and 'drop one character' indexes. The latter stores every combination of each name with single characters dropped; by doing the same for retrieval terms one can obtain a good triage pass before more sophisticated edit distance measures are calculated.

6.2.2 Human processing

Having obtained automatic 'candidate matches', these were then available for human verification via two interfaces. In one the match lists were exported as a spreadsheet for off line processing, which could then be later re-imported; in the other (fig 7), the data was presented in a web interface. Both were showing names from one data set (the source) on the left, the possible matches (targets) on the right, and a computer generated confidence value between. The musicologist could then mark these as 'Y' (yes), 'N' (no) or 'P' (not sure).



InConcert Link Manager

showing 1-50 of

Id	From Id	From Display	?	To Id	To Display	Matches
20393	ABRAMS~	Abrams, <Miss> Harriet	???	1-R-36	Abrams, Harriet	
20394	ALDRICH	Aldrich, <Dr> Henry	???	5-R-57	Aldrich, Henry	
20395	ARNE-M	Arne, Michael	???	13-L-45	Arne, Michael	
20396	ARNE-TA	Arne, <Dr> Thomas Augustine	???	13-R-6	Arne, Thomas Augustine	
20397	ARNE~A	Arne, <Mrs> Ann	???	14-L-9	Arne, Mrs	
20398	ARNE~C	Arne, <Mrs> Cecilia	???	14-L-9	Arne, Mrs	
20399	ARNE~E	Arne, <Mrs> Elizabeth	???	14-L-9	Arne, Mrs	
20400	ARNOLD	Arnold, <Dr> Samuel	???	14-R-20	Arnold, Samuel	

Concerts 1750-1800

Person

abv ABRAMS~
forename <Miss> Harriet
notes <Sen; [one of] the Miss Abrams>. Some might be Theodosia Abrams.

Biography

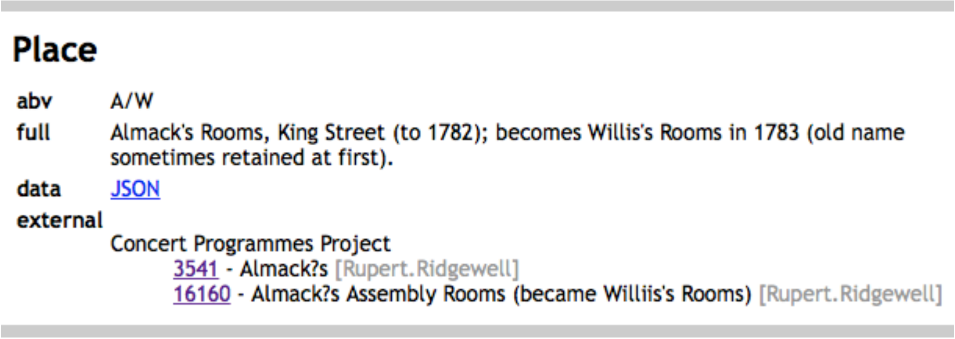
Abrams, Harriet

[E] Abrams, Harriet, soprano vocalist and composer, was born in 1760. She studied under Dr. T. A. Arne, and made her first appearance at Drury Lane in his " May Day,"

Figure 7. Prototype web interface for link checking

In general, verified matches were almost always for the entry with highest automatic confidence score; however, there was no sensible 'critical value' for this confidence score, highlighting the need for human expert evaluation.

The completed spreadsheet or web interaction was processed to create a link dataset listing the connections between the datasets (similar to RDF 'sameAs'). By keeping this separate, it is possible to easily maintain the provenance of the link information, fully automatic or human, and if human by whom (see fig. 8). Different experts may resolve the names in different ways, or decide whether they trust the source of the linkage information (automatic or human) for a particular scholarly purpose. This cross-linking was also used to enable RDF Linked-Data views of the datasets [ND16].



Place

abv A/W
full Almack's Rooms, King Street (to 1782); becomes Willis's Rooms in 1783 (old name sometimes retained at first).
data JSON
external
Concert Programmes Project
3541 - Almack?s [Rupert.Ridgewell]
16160 - Almack?s Assembly Rooms (became Williis's Rooms) [Rupert.Ridgewell]

Figure 8. Links displayed with provenance

Note that while some of this matching was done by the musicologist, some was also performed by the technology partner, who was not an expert. However, the fact that the linking dataset contained provenance: who or what did the matching, made it possible to regard the non-expert's matching as a suggestion, just like the automatic matching. Furthermore, the ability to visualise this provenance (as in fig. 2), means that anyone wishing to make scholarly judgements based on the dataset can take into account the expertise of the matcher.

6.3 Grouping and matching within a dataset

As noted previously, the LC19 dataset of 19th century concert notices/adverts could potentially contain multiple entries relating to the same concert. The remaining (interpretation) phase was to go through these concert notices, work out which ones referred to the same event and create an authoritative entry for each concert. This process the musicologists refer to as 'skewering', but database technologists would think of as entity/object identification or record linkage [Du46, AI07].

This process had acted as a block to progress, as it was so substantial and required expert attention. A major breakthrough was realising that this consisted of (at least) two separable sub-tasks, described earlier: (T6) match – 'skewer' multiple notices referring to the same concert; (T7) merge – combine the data from the notices to create an authoritative record for the concert. It became clear that, while the effort in doing the match task was substantially less than the merge task, still the dataset would become substantially more valuable once the first sub-task was complete.

There is a substantial literature on entity/object identification dating back from the early days of databases [Du46] to semantic web applications [NA12]. Sometimes this involves simple similarity measures such as Jacquard distance between feature sets, or Levenshtein edit distance for string matching. Other researchers have used complex machine learning techniques, including using structural relationships in relational or graph databases [RS06, BG07, GS10]. There is also tool support. OpenRefine (formerly Google Refine) supports the management of data including linking names to entities (possibly more like the name matching in the previous section), although it does not do matching itself, passing this task on to external data services through its Reconciliation Service API [OR18]. RELAIS (REcord Linkage At IStat) is dedicated to the process of record linkage itself [ST15]; it supports a number of different matching algorithms that can be applied to any combination of fields.

However, as with the name matching, because this was part of human–computer process, simpler automatic matching was sufficient combined with methods to make the human task easier. Crucially the matching algorithm was *liberal* in terms of finding potential matches: those that had the same date and similar venue names were matched into groups. This inevitably led to some false negatives (e.g. if the date or venue of a concert changed between notices) and false positives (several concerts at the same venue on the same day). However, the liberal matching was combined with a *conservative* process of marking warnings on those where the match was not almost exact.

This combination meant that it was highly likely that potential matches were already grouped, even if some groups contained more than one event. However the warnings helped to focus attention on groups which might need division by the expert.

A similar process of exporting and importing spreadsheets was used as for the authority file matching, a process that we found extremely efficient in terms of both development time and ease of learning [DC16]. As with the previous interface. the spreadsheet allowed the assessor to attach a level of confidence to the grouping and when the spreadsheet was re-imported, the dataset was updated to include who had performed the group verification.

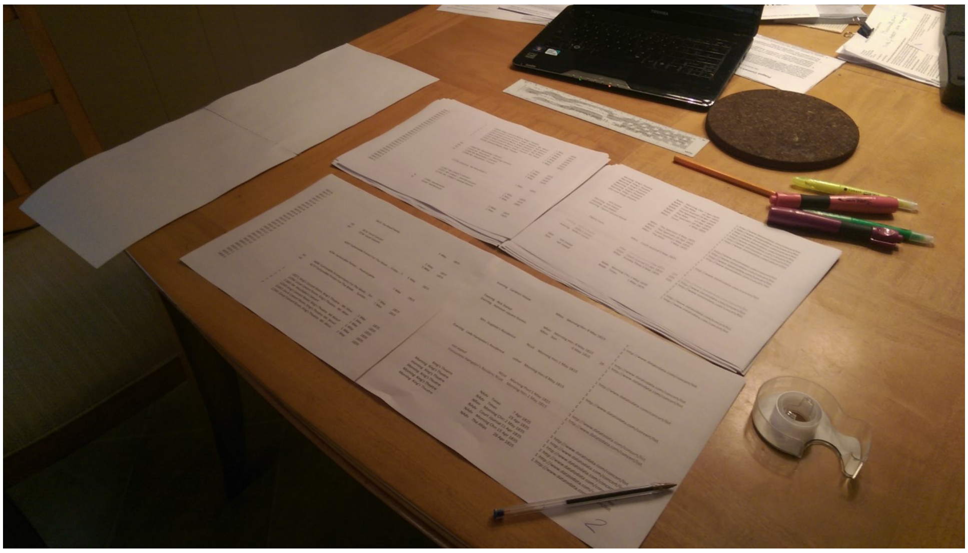


Figure 9. Printed spreadsheet for grouping by hand

7. Discussion – the future for crowdsourcing in digital archives

We saw that while there appear to be many potential tasks suitable for crowdsourcing when preparing a digital archive in the humanities, there are also barriers, especially for macrotasks, which tend to require a level of expertise. However, we have also seen that *In Concert* has employed both automated algorithms and (trusted) non-experts when working on the creation of its datasets.

Based on these experiences, we can revisit the issue of crowdsourcing, looking at the properties of tasks, interfaces and workflows that made it possible to use these non-expert actors. Doing this we see ways in which crowdsourcing by amateurs may be possible within a scholarly culture and identify enabling heuristics.

1. *Understanding values* – Our first and most important step was to understand the scholarly values and academic value systems that drive and constrain scholarly activity. Attitudes that, to an outsider, might seem like academic elitism are in fact rooted in the very real need to maintain a reliable and authoritative corpus. Specific practices may be radically reimagined, but only by understanding and working within a context of deep scholarly values.

2. *Deconstructing tasks* – Macrotasks where scholarly expertise seems essential may be broken down into microtasks, some of which may be amenable to less expert help, with residual less-extensive expert macrotasks. Within *In Concert* this

was highly effective in recognising opportunities for automatic processing, but the same process could identify crowdsourcing potential. Crucially, there is evidence that, where it is possible, decomposing into microtasks increase the quality of results [DT15], which fits well with the scholarly values. Furthermore, the lower volume of the residual macrotasks may make it easier for the scholar to apply contextual understanding.

3. *Deconstructing expertise* – Computer processing may lead to erroneous, weird and occasionally risible outputs, but it is consistent. The trusted non-experts lacked domain knowledge, but were meticulous and (in general terms) scholarly in their approach. Microtask crowdsourcing makes use of very generic low-level skills, such as visual matching. Macrotasks may involve more complex activities, for example, scanning sources for mentions of concerts, but not necessarily the knowledge of the professoriate. Distinguishing types of expertise and skill may help identify places where the 'expert' need not be a domain expert.

4. *Sanity check rules* – These build confidence in the processed data, but also highlight where more expert human intervention is required. In the automated processing this led to updating of rules, or creation of exceptions. In crowdsourced processing this might lead to updating instructions or marking of certain parts of the dataset for more expert processing. Furthermore sanity checking itself may be a human activity, for example the OCR correction workflow in Distributed Proofreaders [DP18] involves multiple human checking stages.

5. *Suggest/confirm workflows* – In both authority file matching and concert notice grouping, the automated matching was seen as creating suggestions for expert confirmation. In the end the experts verified every decision, but for some kinds of tasks scanning work and confirming it can be much faster than doing the task in the first place. Crucially this means that the expert retains control over the final output.

6. *Provenance* – Tracking provenance (who did what and identifying original sources), is of course essential for suggest/confirm workflows, but potentially offers the ability to have datasets with mixed levels of authority. For traditional scholarly work, where the scholar examines individual sources, they can make case-by-case assessment of the extent to which they trust judgements by different individuals in the creation of a digital record. In some cases, if they are uncertain, they can of course check the work by following it back to the sources, a form of just-in-time verification. In more large-scale data or statistical analysis, queries can be formulated to only apply to records with a certain level of verification, or alternatively the query can highlight lists of pertinent unverified records that the scholar can then verify; this is still laborious, but the expert knows that these entries are precisely those needed to address their research question.

8. Summary

In this chapter we have seen how the growing volume of digital material makes crowdsourcing all but essential if humanities research is to keep pace with the burgeoning source material. However, we have also seen that there is a culture clash between the goal of an authoritative reliable corpus and the perceived potential for inaccuracy, inconsistency and unreliability of the amateur.

The easiest approach to dealing with this is to confine crowdsourcing to microtasks that only require day-to-day skills such as visual comparisons. Another approach, more suitable for macrotasks, is to increase the quality and confidence in crowdsourced material, for example, traditional dual keying, sanity check rules, or the multi-stage workflows of Distributed Proofreaders [DP18].

In *In Concert* we adopted elements of both of these, albeit for automatic processing and trusted non-experts rather than fully crowdsourced material. However these were set within a human and digital structure that helped the humanities scholars to retain control of the process. This signposts ways in which crowdsourced material from both microtasks and macrotasks could be similarly included in digital archives so long as their presence is adequately recorded. By making the editorial provenance of data clear, academics can then use their own scholarly judgment as to the reliability of different classes of material and editors for different purposes.

Most crucially, any systems, whether automatic or crowdsourced, need to respect established underlying scholarly values. By so doing we can radically reimagine the processes that lead to the creation of digital archives, but do so in ways that preserve their fundamental academic integrity.

References

[Ac22] Ackerman, Phyllis. (1922). Catalogue of the Retrospective Loan Exhibition of European Tapestries", Taylor and Tayloy, NY. http://www.gutenberg.org/ebooks/57518

[AI07] Ahmed, E., Ipeirotis, P. and Verykios, V. (2007). Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 19 (1):1–16. doi:10.1109/TKDE.2007.9

[vA08] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum, 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science, 321(5895):1465--1468.

[BC00] 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.

[Be04] 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

[BL89] Tim Berners-Lee (1989). Information Management: A Proposal. CERN internal report, March 1989, May 1990. http://info.cern.ch/Proposal.html

[BG07] Bhattacharya, I. and Getoor, L. (2007). Collective entity resolution in relational data. ACM Trans. Knowl. Discov. Data, 1(1):5

[BL12] Bodleian Library (2012/2019). What's the Score at the Bodleian? Bodleian Library. accessed 1/5/2019 http://scores.bodleian.ox.ac.uk

[Bo46] Borges, J. (1946). Del rigor en la ciencia. (tr. 'On Exactitude in Science') Los Anales de Buenos Aires 1.3 (Mar. 1946):53.

[BS97] Brown, J. and Stratton, S. (1897). British Musical Biography: a dictionary of musical artists, authors and composers, born in Britain and its colonies. S.S. Stratton, Birmingham.
OCR text: https://archive.org/details/britishmusicalb00brow
searchable and data version: http://www.datatodata.com/in-concert/BMB/

[DT15] Justin Cheng, Jaime Teevan, Shamsi T. Iqbal, and Michael S. Bernstein. 2015. Break It Down: A Comparison of Macro- and Microtasks. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). ACM, New York, NY, USA, 4061-4064. DOI: 10.1145/2702123.2702146

[CL97] *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).

[CP04] Concert Programmes online database. created 2004–2007. accessed 29/9/2018. http://www.concertprogrammes.org.uk/about/

[CR12] Cowgill, R. and Poriss, H. (eds) (2012). The Arts of the Prima Donna in the Long Nineteenth Century. Oxford University Press.

[DB00] Dix, A., Beale, R. and Wood, A. (2000). Architectures to make Simple Visualisations using Simple Systems. Proc. AVI2000, ACM, pp. 51–60.

[DC14] Alan Dix, Rachel Cowgill, Christina Bashford, Simon McVeigh, and Rupert Ridgewell. 2014. Authority and Judgement in the Digital Archive. In Proceedings of the 1st International Workshop on Digital Libraries for Musicology (DLfM '14). ACM, New York, NY, USA, 1-8. DOI:10.1145/2660168.2660171

[DC16] A. Dix, R. Cowgill, C. Bashford, S. McVeigh and R. Ridgewell (2016). Spreadsheets as User Interfaces. Proc. AVI2016, ACM, pp.192-195. DOI: 10.1145/2909132.2909271

[Dx19] Dix, A. (2019). Creativity – understanding and enhancing technical creativity and innovation. Accessed 11/1/2019. https://alandix.com/creativity/

[DP18] Distributed Proofreaders (2018). Distributed Proofreaders: preserving history one page at a time. Accessed 2/9/2018 https://www.pgdp.net/

[Du46] Dunn, H. (1946). Record Linkage. American Journal of Public Health 36 (12): pp. 1412–1416. doi:10.2105/AJPH.36.12.1412

[FS17] Florian Fink, Klaus U. Schulz, and Uwe Springmann. 2017. Profiling of OCR'ed Historical Texts Revisited. In Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2017). ACM, New York, NY, USA, 61-66. DOI: https://doi.org/10.1145/3078081.3078096

[GS10] Di Gioia, M., Scannapieco, M. and Beneventano, D. (2010). Object Identification across Multiple Sources. Proc. of the Eighteenth Italian Symposium on Advanced Database Systems, SEBD 2010, Rimini, Italy, June 20–23, 2010.

[Go16] Michael Gove (2016) Sky News interview with Faisal Islam, 6 June 2016.

[Gr00] Grove, George, ed.; A Dictionary of Music and Musicians 1450–1889 (1900).

[HA15] Daniel Haas, Jason Ansel, Lydia Gu, and Adam Marcus. 2015. Argonaut: macrotask crowdsourcing for complex data processing. Proc. VLDB Endow. 8, 12 (August 2015), 1642-1653. DOI=http://dx.doi.org/10.14778/2824032.2824062

[IC16] In Concert (2014–2016). accessed 3/1/2016
http://inconcert.datatodata.com

[LT18] Leverhulme Trust (2018). Research Project Grants. Accessed 4/9/2018. https://www.leverhulme.ac.uk/funding/grant-schemes/research-project-grants

[MV92] McVeigh, S. (1992–2014) Calendar of London Concerts 1750–1800. (Dataset) Goldsmiths, University of London. http://research.gold.ac.uk/10342/

[NA12] Nikolov, A., d'Aquin, M., and Motta, E. (2012). Unsupervised learning of link discovery configuration. In Proc. ESWC'12, Springer-Verlag, Berlin, Heidelberg, 119–133. doi: 10.1007/978-3-642-30284-8_15

[ND16] T. Nurmikko-Fuller, A. Dix, D. M. Weigl, and K. R. Page (2016) In Collaboration with In Concert: Reflecting a Digital Library as Linked Data for Performance Ephemera. In Proceedings of the 3rd International workshop on Digital Libraries for Musicology (DLfM 2016). ACM, New York, NY, USA, 17-24. DOI: 10.1145/2970044.2970049

[OR18] OpenRefine: Reconciliation Service API. (accessed 24/9/2018).https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API

[REF12] 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/

[RS06] Rendle, S. and Schmidt-Thieme. L. (2006). Object identification with constraints. Data Mining, 2006., 1026–1031, http://www.ismll.uni-hildesheim.de/pub/pdfs/Rendle_SchmidtThieme2006-Object_Identification_with_Constraints.pdf

[Ri07] Chris Rusbridge (2007). Arts and Humanities Data Service decision. DCC News, 6 June, 2007. Digital Curation Centre. http://www.dcc.ac.uk/news/arts-and-humanities-data-service-decision

[ST15] Scannapieco, M., Tosco, L., Valentino, L., Mancini, L., Cibella, N., Tuoto T. and Fortini, M. (2015). Relais User's Guide – Version 3.0. Technical Report, Italian National Institute of Statistics (Istat). July 2015, doi:10.13140/RG.2.1.1332.5922

[SL18] Heinz Schmitz and Ioanna Lykourentzou. 2018. Online Sequencing of Non-Decomposable Macrotasks in Expert Crowdsourcing. Trans. Soc. Comput. 1, 1, Article 1 (January 2018), 33 pages. DOI: https://doi.org/10.1145/3140459

[TM16] Transforming Musicology (accessed 3/1/2016). http://www.transforming-musicology.org

[VG14] Vobl, Thorsten, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz. 2014. "PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts." In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 57–61. DATeCH '14. New York, NY, USA: ACM. doi:http://doi.org/10.1145/2595188.2595197.

[Wi19] Wikipedia (2109) Arts and Humanities Data Service. Accessed 1/5/2019 https://en.wikipedia.org/wiki/Arts_and_Humanities_Data_Service