http://dx.doi.org/10.14236/ewic/HCI2022.2

Tools and technology to support rich community heritage

Alan Dix
Computational Foundry
Swansea University, Swansea, Wales, UK
www.alandix.com
alan@hcibook.com

Troedrhiwfuwch Team
Elizabeth Jones[1, 2], Carys-Ann Neads[2], Vince Davies[2]
[1]Swansea University, UK
[2]Troedrhiwfuwch Community, UK
whereweare.org/troedrhiwfuwch

InterMusE Team
Rachel Cowgill[3], Charlotte Armstrong[3], Rupert Ridgewell[4],
Michael Twidale[5], Stephen Downie[5], Maureen Reagan[5], Christina Bashford[5],
[3]University of York, UK; [4]British Library UK; [5]University of Illinois Urbana-Champagne, USA
intermuse.datatodata.org

This paper explores ways in which scholarly skill and expertise might be embodied in tools and sustainable practices that enable communities to create and manage their own digital archives. We focus particularly on tools and practices related to the recording and annotation of digitised materials. The paper is based on co-production practice in two very different kinds of community. Although the communities are different we find that tools designed specifically for one are valuable for others, thus offering the promise of general tools to support community-centred digitisation and potentially also traditional archival practice.

community heritage, digital archives, digital storytelling, democratising digitisation

1. INTRODUCTION

It is often said that Covid-19 has awakened many to the importance of community. However, it also seems that communities are under threat, the sense of identity and belonging drowned in the homogenisation of global media and rootlessness of modern living. The heritage, culture and history of communities is one of the things that nurtures this sense of belonging, and yet is also precarious. When flood or fire destroy some part of a national museum or art gallery it makes headline news, and yet every day shoe boxes of memorabilia or old papers are discarded or lost as people move, die or downsize.

One of the bulwarks against loss of large collections is digitisation, which also opens the archive to more widespread scholarly study of materials and public dissemination. Furthermore, in the internet age, digital resources offer visibility and thus influence. The role of the latter was particularly crucial during periods of Covid lockdown (Vayanou et al. (2020)), but also has the potential to allow materials to be re-presented in ways that reach audiences who would not usually visit cultural institutions.

Can these benefits of digitisation be harnessed for small communities, offering them the means to preserve, explore and publicise their own heritage and stories? More crucially, can we *democratise digitisation* – make it available, not simply when a team of university researchers parachute in to offer expertise and resources that are necessarily limited in time and scope, but embody that skill and expertise in tools and sustainable practices that enable communities to manage their own digital archives?

In this paper we explore some of the questions around these issues and also present early prototypes that we hope will be valuable across different kinds of communities and settings. In particular, we will focus on tools and practices related to the recording and annotation of digitised materials – that is, the creation and management of digital community archives. We will also see that the boundaries between collection, curation and communication are far more fluid in community-centred digitisation than in traditional archival practice ... even though the latter is itself at a point of flux (Dix et al. (2014); Hoyle (2022)).

We focus on two very different kinds of community, both are *pseudo-geographic* in that they have elements of physical locality, but include members not defined simply by where they live. One is a small village, Troedrhiwfuwch, in the Welsh (ex)coal-mining valleys, that was physically demolished in the 1980s but where the community still lives on. The others are a group of local music societies in Belfast, Huddersfield and York in the UK, all of which were part of a post WWI initiative to use music to foster a sense of international connection.

While the social demographics and reasons for existing are very different, we will see that there are commonalities; in particular, prototypes designed for each have value for the other. This suggests that bespoke development and rich co-design for specific communities can lead to tools and processes useful to many.

In the next section we will review some of the conceptualisations of 'community', which is critical in so many disciplines from human geography and social science to health, as well as heritage. We then look at the two communities we are studying: Troedrhiwfuwch (Section 3) and the extant regional branches of the organisation founded as 'the British Music Society' in 1918 (Section 4). For each we will first describe the community, the engagement between researchers and community members and initial concepts and themes emerging from them; we will then look at a prototype designed for the community: TalkOver for Troedrhiwfuwch and OcrMarkup for the music societies. After describing each community and its prototype, in Section 5, we will look at what happened when the communities were exposed to the prototypes for the other community and lessons we can take away from this.

2. DIMENSIONS OF COMMUNITY

Community is a word we all recognise and yet almost certainly all understand in different ways. Most readers will be academics and in parallel be part of a local community around their home, maybe a separate 'home' community where they were brought up, a university or departmental community, including academics, administrators and students, and a professional community, for example as HCI researchers.

The AHRC Connected Communities programme in the UK (itself a community of practice of researchers of 'community') produced a number of detailed reviews and commentaries, which together capture some of the complexities of community (Crow and Mah (2012); Studdert and Walkerdine (2016)). This includes the way 'community' emerged as subject of study in the 19th Century, largely as something being lost (Walkerdine and Studdert (2012)), and highlights that 'community', despite a large literature and being the focus of many government initiatives, is still often poorly defined – a 'spray-on term' (Walkerdine and Studdert (2012)), or 'slippery concept' (Craig and Mayo (2011)) subject to 'disciplinary confusion' (Studdert and Walkerdine (2016)).

The most obvious concept of community is geographic – people in a village, town, or urban neighbourhood, the idea of one's own *campanilismo* (bell tower) in Italy, or *milltir sgwâr* (square mile) in Wales. However, researchers in human–computer interaction will also be familiar with the anthropological concept of communities of practice (Lave and Wenger (1991); Wenger (1999)) which are often linked to professions, or other forms of interest group.

This distinction is fluid and many communities are *pseudo-geographic*, being associated with a place, but not necessarily living in that place (e.g. university alumni), or residing in a smaller/larger space, but based around interests or characteristics in common but not necessarily shared by everyone in the region (e.g. religious or ethnic communities within an area, or chambers of commerce). The communities we will describe below are both pseudo-geographic — one dispersed, but linked by a single common physical origin, the other based around common musical interest within a wide geographic area.

As well as these dimensions of place and interest (Willmott (1986)), many conceptualisations look more at what a community does, or how it is experienced including a sense of identity (Willmott (1986)), connection, difference, boundaries and development (Crow and Mah (2012)), or the action of communing, relationality and sociality (Studdert and Walkerdine (2016)).

These characteristics of community bridge geographic and thematic dimensions and emphasise the shared aspects of communities of many kinds. It is therefore not so surprising that we shall find that tools created for one kind of community end up being applicable to others.

3. PEOPLE OF A LOST LAND

3.1. Context

Troedrhiwfuwch was a small coal-mining village nestling in the eastern slopes of the Rhymney Valley in South Wales. From 94 households, 110 young men left for the First World War, 21 of whom never returned. This was one of the greatest concentrations of war service enlistment in the

Figure 1: *Troedrhiwfuwch before and after evacuation.*

country for the size of the small community, which totalled 600 – a commitment and sacrifice recognised by King George V. Then in 1976, the village was condemned. In 1966, 28 adults and 116 children lost their lives in Aberfan, another mining community, when a rain-soaked coal tip, the discarded rocks and coal dust from deep mining, slid down the mountainside and buried the village school. In the aftermath, surveys assessed the stability of other mountain sides and coal tips across the area. The mountain above Troedrhiwfuwch was deemed at risk, and, over a number of years, the people of the village were rehoused. Most of the village structure was demolished by 1985. Today, only two houses and the war memorial and garden remain as a sign of the place that once was (Figure 1).

However, the diaspora of Troedrhiwfuwch, or 'Troedy', as it is known locally, have not forgotten their past. Each year on Armistice Sunday, a group congregates at the war memorial for an act of remembrance and there is an active Facebook memories group. A smaller group is also active, gathering photographs and documents from local people and scouring national archives for material connected to the village. This includes a digital archive of more than 1,400 items and extensive paper material. A particular focus has been on the First World War, especially following 2014–2018 centenary events, and given the War Memorial and the adjacent Memorial Garden (on the site of the demolished church) are some of the few remaining signs on the ground.

This diaspora community is not just the old who lived their life there, although some are in this category. Many only know of the village through trips as children, when parents and grandparents would point to a patch of grass and tell them stories of the place where an aunt or cousin once lived. Some lived in or visited the village as a small child while it still stood, but one of the most active members of the history group was born well after the long terraces, which once lined the roadside, were demolished.

3.2. Engagement

One of the authors works for a university as well as being a family member of the Troedrhiwfuwch community. She acted as the first point of contact. Since March 2021 a small group of academic researchers and community members have met, largely informally, around a dozen times. This has been mostly using video conferencing, but there have also been several site visits, albeit limited due to Covid. The latter included walking the ground of the village itself, and also visiting a church in a neighbouring village where the interior furnishings of the demolished Troedrhiwfuwch Church have been used to create a small side-chapel, forming a compact reproduction.

As with any co-production exercise, the team wanted to embed the principles of equality, diversity, accessibility and reciprocity in putting co-production into action (Social Care Institute for Excellence (2015)), and there was a period of mutual enculturation. On the one side, the non-Troedrhiwfuwch academics built an understanding of what it means to be part of the community. This was accomplished principally through *story-telling* often focused around digital artefacts, or walking the ground itself. On the other side, the community members built an understanding of the potential of digital technology to help them preserve, organise and disseminate their heritage materials. This was facilitated by the production of early envisionments using PowerPoint scenarios and paper-and-card low-fidelity prototypes.

3.3. Emerging concepts

One way to view engagement between university researchers and community members would be as an expert–amateur or expert–end-user conversation, based on mutual respect but with different roles. There is truth in this; however it misses the rich and diverse ***expertise of the community*** members themselves. There are obvious elements to this expertise: personal knowledge of events through direct experience or conversations with others – connection points into human networks and understanding of the needs and aspirations of the community. However, this is only a part of the story. The Troedrhiwfuwch archivists have a knowledge of historic sources such as military records, genealogical resources and census reports. This facility with primary resources is complemented by a synthesised knowledge of the historic relations between people and events, similar to that which the (non-historian) academic members of the team have observed in their academic colleagues' historical knowledge. This is not to equate academic and community historical expertise and approaches, but to problematise words such as 'amateur' and 'expert'.

One of the key differences is that community history is often ***intimately connected to family history***.

The people in a photo are not simply objects of study, but great-aunts and grandparents, with stories that are part of one's own story. Equally, these ***personal stories are often universal stories***, and (for those outside of the community) the stories of individuals to whom one has no personal connection can not merely fascinate as stories, but can parallel one's own experience – lessons not lost on the producers of popular TV family history programmes.

We have noted the extensive nature of the existing digital archive including photographs, documents, census records. However, whilst the individual items are preserved and organised, the meta-data, the knowledge of what things are and how they relate to one another, is largely in the heads of the community archivists. This includes the provenance of items – who donated a photograph or pamphlet and from which website or military archive an item was downloaded. This is important from a scholarly viewpoint, but also practically – for example, if items are presented externally on a community website are there IP restrictions on images? Filesystem design has hardly changed since the 1970s – each file is isolated and related to others only by their location in the folder/directory hierarchy. Archivists, both professional and lay, need better ways to ***annotate and connect***.

Expert knowledge is often tacit – only brought to bear in particular circumstances and contexts. This is equally true of community knowledge – ***people, places and artefacts*** elicit knowledge and stories. One example of this was seen while walking the ground of the village. The precise position on the ground of demolished houses are often half guessed in relation to natural outcrops of rock. Then one of the community members said, "my house was here". The house had gone, but the drain in the road had lain by the outer corner of the house and the drain remained.

There is a ***fragility and precarity*** to these memories. This is true of the personal memories of ageing people, but also for physical artefacts. Troedrhiwfuwch emphasises that even buildings and solid rock may shift or fall. Between memories and masonry are many photographs, small items and documents that live on mantelpieces, or in attics. When a person dies, not only are their memories lost, but these objects, embodying community heritage as well as personal significance, may end up on the fire, or in a junk shop or skip.

3.4. Prototype: Talkover – capturing stories about photographs

TalkOver is an experimental web app that makes it easy to record stories about pictures. It can be used for gathering oral history about old photographs or documents, or for any application where you want to produce narratives about images.

This demo arose originally from experiences during meetings between the Troedrhiwfuwch community and researchers. The extensive archive of photographs and documents is impressive in itself. However, as soon as any one of the photos is opened, community members start to tell stories: some about past relatives that they were told as children, some from research they have done in other archives or war records. The details that make the photographs come to life and connect them together are in the heads and memories of the community, but not recorded in their digital archive.

Narrative and storytelling have always been an essential part of community history. The Oral History Society (2022) cites examples from as far back as the 8th Century and Sharpless (2008) looks back to Heroditus in the 5th Century BCE. This accelerated in the 19th Century, especially in relation to folk tales and songs. However, the emergence of audio recording and especially magnetic-tape recording created the modern field of oral history. Digital technology has further transformed the collection and curation of audio material (Lambert and Frisch (2019)) including geo-coding stories whilst walking (Zembrzycki (2013)). In presentations of oral history for public access, the spoken word is often illustrated in professionally edited multi-media presentations, with the voice overlaying still images. However, we wanted something similar, but with the ease of pointing at people in a photograph as one does when one sits side-by-side.

TalkOver addresses this by recording not just the speaker's voice, but also by allowing the person being recorded to point at a digital image using either their finger on a tablet or mouse on a laptop screen. As the user touches the picture, a small halo temporarily appears at the point they touched as feedback (see Fig. 2). The locations are recorded together with their time-stamps. The audio and marks are stored alongside the image and can then be replayed. This creates a rich playback akin to a crafted multi-media presentation, but with the immediacy of a side-by-side telling. As the work was performed during Covid lockdown, this was especially poignant.

As well as offering a rich form of collection, the marks associate areas of the image with points in the story. If faces or objects in the images are also indexed, automatically or by hand, with people and themes, then this offers the potential for interlinking semantic annotations and continuous media.

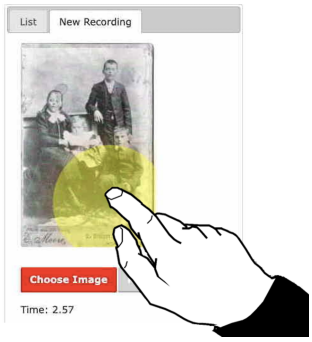

Figure 2: *Talkover recording in progress.*

There is an art to interviewing for oral history, and the system does not replace that. However, the act of talking about something is very natural and thus offers a way for less skilled interviewers to collect oral history, as well as offering an additional tool for the professional oral historian.

3.5. Under the Hood

TalkOver is built as a standalone web app – that is, all processing and storage are local to the user's machine. This allows sharing of usable prototypes, without complex installation and without the need for extensive cloud or server infrastructure.

New images can be added by drag-and-drop or the file chooser, using standard cross-browser W3C file APIs. WebAudioRecorder.js is used for audio capture, which is built on the W3C WebAudio API. This performs all recording and encoding in Web Workers in the browser, meaning that no external transcoding is needed. The audio is stored in the browser's IndexDB store, which can accept large media data and provides persistent local storage. This is used both to store the raw media (images and audio) and also the data structures describing the user's pointing actions (essentially time-stamped coordinates).

Following the web technology theme, the import/export format for backing up and sharing TalkOver recordings is simply an HTML file (see Figure 3). This includes the complete image and audio media base64-encoded in Javascript variables, as well as further meta-information in sections demarcated by easily identifiable comments. These TalkOver HTML archives can be loaded back into the TalkOver application, which parses the HTML and extracts the media variables. A GUID is generated for each new TalkOver recording and stored in the HTML archive format, so that if a backup is reloaded or a shared recording loaded twice it can be connected to the original recording.

```
<!DOCTYPE html>
<html>
<!-- @@wrappeddata       -->
<!-- @version: 0.1       -->
<!-- @type: talkover     -->
<head>
    <meta charset="UTF-8">
    <title>Talkover</title>
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
</head>
<body>
    <h1>TalkOver - Playback</h1>
   <div id="playback"></div>
    <br  clear="both">
     <script>
         // @!meta
         var meta_info = { "type":"talkover" };
         // @!end meta
         // @!data
         var talkover_info = {"version":"0.1","creation_time":"2022-01-16T19:35:19.113Z","au
         var image_url = "http://whereweare.org/test/record/images/Hillman-family.jpg";
         var audio_url = "data:audio/wav;base64,UklGRiTACQBXQVZFZm10IBAAAAABAAIARKwAABCxAgAI
         // @!end data
     </script>
     <script src="https://datatodata.org/talkover/demo/v001/js/bootstrap_v0.js"></script>
</body>
</html>
```

Figure 3: *TalkOver export format as HTML; note the base64-encoded* `audio_url` *is typically between 5 and 50 million characters long.*

The HTML content is minimal, but includes a link to a single JavaScript bootstrap file, which allows smaller recordings to be opened by double-clicking the HTML file without explicitly importing. This is similar to the self-describing '#!' prefix for running script files in Unix. As they encapsulate all the media, these HTML archive files are large (around 120Mb for a ten-minute recording), but recordings of up to 2 minutes have been 'click opened' in Safari and up to 5 minutes in Chrome (both in MacOs). We have not yet hit the limit for import/export sizes, and so, as it is intended to be for relatively short recordings, the format seems sufficient for this purpose.

4. REGIONAL MUSIC SOCIETIES

4.1. Context

The British Music Society (BMS) was established in 1918 to restore international collaboration and exchange between British and overseas musicians after the twin catastrophes of the Spanish Influenza and World War One, and to empower amateur musicians in organising and promoting their own concert series. The BMS was formed by the progressive musical author, educator and organist Arthur Eaglefield Hull (1876–1928), with chapters in towns and cities throughout the UK and beyond. Many of these chapters remain active today, and while they may have limited knowledge of their shared origins in Hull's BMS, they have amassed significant archives over the past century that shed light on the rich history of this extraordinary initiative and the broader role of music in regional community life. The Internet of Musical Events: Digital Scholarship, Community, and the Archiving of Performance (InterMusE) is a two-year project funded by the AHRC's UK-US New Directions for Digital Scholarship in Cultural Institutions scheme (Ref. AH/V009664/1). The project brings together a team of scholars from humanities and computing backgrounds to work with three former chapters of

the BMS: the Belfast Music Society (BeMS), British Music Society of York (BMSY) and Huddersfield Music Society (HMS). These institutions are eager to take stock of their histories and document their collections, and InterMusE is working with them to capture and link different forms of data relating to musical events with a view to creating a dynamic, open-access digital archive of musical ephemera.

The collections of the BeMS, BMSY and HMS comprise diverse material types, from concert programmes and other performance ephemera, to newspaper reviews and administrative records. In each case, some physical materials are stored in local archives or libraries, while others are kept in society offices and private homes. As such, the materials have undergone varying degrees of cataloguing, digitisation and preservation. Each society has representative members, volunteers or employees, who have taken a keen interest in its archival collection. Drawing on a range of professional, self-taught and instinctual knowledge, these representatives – the custodians of the collections – have taken steps to ensure the preservation of their society's archival materials for future generations. By working with them to capture and link the data from these materials, as part of a digital archive, we aim to improve access to the archival collections and empower society members to explore and engage with their rich histories. We are also exploring ways in which the expertise of these community members can be used to incrementally enrich the historical records of these societies. The digitised materials will be enhanced with item descriptions and transcriptions, personal recollections and oral histories.

4.2. Engagement

From university-based researchers, archivists and programmers, to citizen researchers, amateur musicians and fans, InterMusE brings together a range of different stakeholders. The project places a strong emphasis on collaboration and co-production with these societies and their communities, and resists privileging any one stakeholder group over any other. To ensure that the digital archive produced is both a valuable research resource and fit-for-purpose for the societies, the approach has been shaped by a desire to design and create a digital archive with (rather than for) the societies and their communities.

One of the first steps was to take stock of the current collection and preservation activities in each society and understand various stakeholders' visions for the project. These collection assessments were conducted in April 2021 as informal, unstructured interviews. This kind of informal interaction proved effective in establishing a foundation for trust and reciprocal exchange between the project investigators and citizen groups. In July 2021, a second set of group information sessions, also conducted over zoom, provided a forum for society members to share their thoughts on the project and voice any questions or areas of concern.

4.3. Emerging concepts

Several of the themes that arose in Troedrhiwfuwch have parallels in the music societies.

The ***expertise of the communities*** was again very evident. Some of this is in terms of skills and experience brought into their roles; for example, the music-society committees include several who have retired from senior roles in public service and industry. In addition, one member has developed a complete database of concerts including itemisation of the programmes.

The interweaving of ***community with personal and family history*** is also evident, although in a different way. Committee members are often long standing, so that members looking through old committee minutes or concert programmes see names of current and past friends and family. In addition, the concert venues mentioned in early 20th-century programmes are typically in local places that in many cases are still standing and may even be recent venues. That is, ***people, places and artefacts*** elicit knowledge and stories in a very similar way.

Early investigations by humanities students in Illinois have also uncovered many connections between performers in the UK societies and the classical musical scenes of the USA and continental Europe, speaking to the ***universal*** nature of local history within the global community of interest – exactly Arthur Eaglefield Hull's vision. Although elites in London may refer in publications to "the provinces" we find that people with national and international reputations often travelled to places like Huddersfield to perform. By connecting information in the digitised concert programmes to other databases we can see richer connections with larger social and political impacts, such as the Russian Revolution and the disintegration of the Austro-Hungarian Empire and how that influenced who performed what in British towns in the 1920s.

Issues of ***fragility and precarity*** are also common in discussions. In one case a member of one of the societies, who had an extensive collection, died and the documents could easily have been lost. Happily, their spouse knew about them and the passions for preservation that lay behind them, and passed them to a current committee member. However, it was

Figure 4: *OcrMarkup showing areas marked on screen and annotation fields.*

evident that this was a moment when crucial records might have been lost for ever.

In the Troedrhiwfuwch archive many of the textual items (e.g. war records) are already digital, and many of the more internal community artefacts are photographic and visual. In contrast, the music societies have large paper repositories of largely textual content, such as concert programmes, reviews and meeting minutes. The immediate need is to ***digitise and then catalogue relatively structured information*** from them. That is, while the need to ***annotate and connect*** is present, it is of a more structured form, even though the individual formatting of that information differs markedly from programme to programme.

4.4. Prototype: OcrMarkup – from text to meaning

This envisionment prototype was created to show how OCR can be used to help add semantic markup to scanned documents. This is specifically for situations where a level of expert judgement is important: a fully automated solution is not appropriate, but it is important to make the most of what the computer can do to help.

This prototype arose directly from early discussions in the InterMusE project where we are working with concert programmes. Commonly the output of OCR is a continuous text, sometimes with attempts to deal with common forms of document structure, such as columns. The text versions of documents in Project Gutenberg or HathiTrust archives are good examples of this. This works well for linear text, such as a novel, but less so for structured documents.

In previous projects we had created digital versions of two 19th-century catalogue-style documents, the British Musical Biography (Brown and Stratton (1897)) and Gazetteer of Scotland (Wilson (1882)). These were semi-structured, although care was still needed to identify entry headings (personal names or place names) semi-automatically, for example, using all-caps.

Concert programmes are far more complex, with multiple sections for performers, dates and times, pieces played, etc., and the layout of these often differs even within a single concert series (see Figure 4, left). Furthermore a substantial portion of the text consists of personal names and titles of pieces (in a variety of languages), making automatic processing difficult. For example, off-the-shelf OCR might take a column of performer names and concatenate it into a single unpunctuated paragraph:

```
ADOLFO BETTI ALFRED POCHON NICOLAS MOLDAVAN
IVAN D'ARCHAMBEAU
```

The variety of concert-programme structures means that **human-intensive intervention** is essential in order to extract meaningful semantics. Happily, for community-based digitisation that human-intensive intervention is possible. However, we also want to make as much use as possible of OCR in order to make the human task as fluid as possible.

While the final version of OCR is often a linear text, earlier stages of the OCR pipeline retain the precise location on the page of each character, word or phrase. Google Vision API was used initially for OCR extraction, but the current prototype uses Tesseract.js if there is no existing markup. The latter occasionally misses words that are recognised by the Google cloud service; however the differences are marginal and for community use the advantages of open source and a free-at-point-of-use service outweigh the slightly better quality of Google Vision. In later versions we plan to allow configuration of OCR services, including use of OCR embedded in PDF when available.

The OcrMarkup prototype allows the user to select and name areas of the image and automatically extracts the OCR text for the region. Figure 4 shows this in action. The user has dragged out a series of areas in the image and then for each region, as it is selected, the text for that region is placed in a corresponding area in the right-hand column. The user has then labelled these areas “venue”, “date”, “time”, “title” and is in the process of typing “performers” for the most recently identified section. If the user resizes the section on the image, the text in the named annotation is automatically adjusted.

On its own OCR is useful, to allow free-text searching of large digitised collections. It is also possible to automatically identify common types of data, such as dates or people’s names. However, when a human looks at a document they can identify more detailed and specific areas, such as the title of a concert,

or who was performing, creating a rich semantics for each document. While this human-in-the-loop identification of areas is a simple technique, the only other system of which we are aware offering such a facility is Lace0.5 (Robertson (2021)). However, due to a different use domain (semantic markup of the Open Greek and Latin corpus), it adopts a fixed vocabulary for marked sections rather than the open annotations allowed in OcrMarkup.

Lambert and Frisch (2019) describe their transition from linear models of content curation to a hub model, where a core of raw data (e.g. recordings or photographs) gives rise to numerous smaller or larger collections of 'cooked' data, interpreted and annotated by different tools for different purposes and audiences. Our own work also emphasises these more incremental approaches, layering different interpretations and processing, automatic and human, by scholar or community (Dix et al. (2014); Armstrong et al. (2021)).

TalkOver fits into this broad process. Annotations are added incrementally based on the purpose and goals of the user. For example, when a programme is first scanned a community archivist may want to simply annotate key features such as the date, venue and title of the concert, in order to create a bare-bones listing of events. Later another community member might be looking for references to a particular family of musicians, use free-text search to find candidate documents and then mark-up relevant parts. Each person's efforts add to an evolving semantically annotated digital archive.

4.5. Under the Hood

Like TalkOver, OcrMarkup is built as a standalone web app for ease of distribution.

As noted, the first prototype used Google Vision API, but the current version uses Tesseract.js as this executes within the browser (asynchronously as a Web Worker). However, a wrapper class makes the annotation code independent of the choice of OCR engine and, where present, Google Vision OCR can be used. OcrMarkup uses word-level OCR and ignores larger phrase/line structures provided by Google or Tesseract, as text has to be re-threaded within selected regions. Instead a simple custom algorithm is used to detect co-linear text.

OcrMarkup shares the same HTML framework as TalkOver for import/export of completed OcrMarkup annotation and pictorial browsing of past annotations. In both OcrMarkup and TalkOver, MD5 digests are calculated for all immutable media to make it easier to connect multiple annotations to the same underlying image.

5. DISCUSSION – SHARED VALUE

The two prototypes described here were designed and attuned to the specific contexts of the different communities. There are some common features, notably both are pseudo-geographic – they are associated with specific places, but the people do not live alongside one another. This means that community communication and coherence is through specific events and online means such as Facebook. However, the Troedrhewfuwch community do have an identifiable, albeit uninhabited, patch of ground, whereas the music societies are intrinsically dispersed, and have always have been so.

The groups differ markedly socio-economically, and more fundamentally in purpose. For the music societies their history is an essential part of their identity, but in the end it is secondary to their ongoing musical passion. For the Troedrhiwfuwch community, history and heritage are central to their activities and goals. Correspondingly the prototypes that arose from the two groups are very different. We can think of various stages of heritage archives: **collecting** primary and secondary material, **curating and organising** this to enable future use, and finally **communicating** within and beyond the community. Both prototypes are focused on the first of these, collecting, but have a different tenor: TalkOver is focused on informal reminiscence, whilst OcrMarkup is more clearly archival in nature, reflecting the differing purposes and backgrounds of the communities and the co-production activities that gave rise to the bespoke designs.

The surprise, that perhaps should not have been a surprise, is what happened when each prototype was demonstrated to the other group.

When TalkOver was shown to the InterMusE academic team they immediately saw potential value and it was included in an upcoming meeting with music-society members. This was a very early version of TalkOver and it was hard to change the image used, so the demonstration was with a photograph of people from Troedrhiwfuwch (Figure 2). Despite the unfamiliar material, the music-society members also instantly saw potential applications, thinking particularly of long-standing members of the music society who could talk about old concert programmes or AGM minutes adding anecdotes, identifying people, and more.

Following this, the OcrMarkup demo (again in early form) was presented to the Troedrhiwfuwch community. The document was the concert programme in Figure 4, not a local document. This was partly due to the difficulty of changing the document as at that stage the document was being passed by

hand through the web portal of Google Vision API. However, it was also less obvious how it would apply, as many important documents, such as census records or birth certificates, were hand-written. Perhaps because of this, there was not a similar "aha" moment to that when TalkOver was demonstrated to the music societies.

However, a few months after this the Troedrhiwfuwch community approached the research team to ask if the OcrMarkup application was available for use. A new and important document had been added to the community archive and they realised that this was the perfect tool to use for that.

In each case, the 'bespoke' tool custom-designed for the specific needs of a particular community turns out also to be of use to the other very different community. As noted, this perhaps should not have been surprising. Studies of 'single person design', where an application has been targeted at a single individual, found that even the most personalised application was appreciated by others (Razak (2008)); indeed many successful web applications have arisen out of such situations, Wordle being perhaps the most recent example. Similarly, there are enough deep commonalities between apparently different communities that solutions targeted at one are of value to others.

This is very encouraging. There are many projects where universities have worked closely with community groups to create innovative prototypes for community heritage and communication (Taylor and Cheverst (2009); Dix et al. (2016); Beel et al. (2017)). However, if we really want to democratise digitisation, to put tools for digital heritage into the hands of communities, we need to create *reusable* tools. While this at first seems at odds with co-production, in fact our experience is that the creation of applications to help specific situations and the design of tools for general use can go hand in hand.

This is not to say that every tool designed for a specific community will be useful for all others, but that for each targeted tool, there will be a number of other communities for which it is also a useful or even ideal solution. This has been explored at an individual level in *designing for peak experience* (Dix (2010)), which highlights the difference between 'good enough for all' designs, for universal use such as a word processor, compared to 'best for some' applications such as game design. For these 'peak experience' applications a viable, and often the best, development path is to optimise for an individual and only when it is right for that person to attempt to generalise for a slightly larger group. We suggest that this is also a viable, and maybe the preferable, development path for communities also.

6. CONCLUSION AND FUTURE WORK

There is a temptation to try and develop a one-shot one-size-fits-all application, including when designing an archival database. This may be reasonable for large-scale institutions where procedures can be formulated and staff can be trained in particular practices and formats. However, for local communities, we must design for the needs and peculiarities of each. We may not know what people want to say about digitised artefacts until we give people the opportunity to tell us – and that in turn depends on widening the range of who gets to tell us things.

We also need to design for the unexpected. We may not know what we will find in the archive until we have finished digitizing – as was seen in the delayed realisation of the potential of OcrMarkup by the Troedrhiwfuwch community. That means that as well as designing for particular needs right now, we also need to design for ease of revision (refactorability), to make it feasible and affordable to redesign to accommodate future needs and use scenarios. In general, we may not know how the database of digitised artefacts will get used in the future, or how and why it might get interconnected with myriad other databases with all kinds of different content. This underlies our own use of flexible semantics and annotation, but we are aware that we need to also find easy ways to modify these and/or connect them to external ontologies and authority files.

Looking at the two prototypes, while these are developed for different needs and purposes, they also share common features. Both are focused on annotation of images: one linking pictorial/photographic images to added audio commentary, and the other linking textual areas to named attributes. If the same programme were semantically annotated and also had TalkOver stories, we may want to be able to search annotations by name and then use this to index stories that point to the faces of the named people. In some ways this is rather like facets of an underlying semantic model. One could create a general 'do it all' application for media annotation, or indeed select one that already exists, but that would lose the specific qualities and simplicity that make each tool work. Our challenge is to find ways to have multiple targeted applications that share sufficient common data representation to enable sharing and linking, yet are still flexible enough to make entirely new co-produced applications possible.

We are doing all of this in the context of community heritage and, more widely, historical archives. However, we are also aware that many of the issues we face in looking at this larger picture of connection, curation and annotation are shared in other domains, for example data analysis. We hope that by keeping

focused on our domain, we also create concepts and solutions that may be useful more widely – just as the communities we have described here found uses for the tools designed for each other.

More information on the projects and prototypes described here at: `https://www.alandix.com/academic/papers/BHCI2022-community/`.

REFERENCES

Armstrong, C., R. Cowgill, A. Dix, C. Bashford, J. S. Downie, M. Twidale, M. Reagan, and R. Ridgewell (2021). Towards a foundation for collaborative digital archiving with local concert-giving organisations. In *8th Intl Conf. on Digital Libraries for Musicology*, pp. 41–49.

Beel, D., C. Wallace, G. Webster, H. Nguyen, E. Tait, M. Macleod, and C. Mellish (2017). Cultural resilience: The production of rural community heritage, digital archives and the role of volunteers. *Journal of rural studies 54*, 459–468.

Brown, J. D. and S. S. Stratton (1897). *British musical biography: a dictionary of musical artists, authors, and composers born in Britain and its colonies*. SS Stratton. Digital version 2016. `http://datatodata.org/in-concert/BMB/`.

Craig, G. and M. Mayo (2011). *The community development reader: History, themes and issues*. Policy Press.

Crow, G. and A. Mah (2012, 3). Conceptualisations and meanings of 'community': the theory and operationalization of a contested concept. Technical report, University of Southampton and University of Warwick. Report to the Arts and Humanities Research Council.

Dix, A. (2010). Human–computer interaction: A stable discipline, a nascent science, and the growth of the long tail. *Interacting with computers 22*(1), 13–27.

Dix, A., R. Cowgill, C. Bashford, S. McVeigh, and R. Ridgewell (2014). Authority and judgement in the digital archive. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pp. 1–8.

Dix, A., A. Malizia, T. Turchi, S. Gill, G. Loudon, R. Morris, A. Chamberlain, and A. Bellucci (2016). Rich digital collaborations in a small rural community. In *Collaboration Meets Interactive Spaces*, pp. 463–483. Springer.

Hoyle, V. (2022). *The Remaking of Archival Values*. Routledge.

Lambert, D. and M. Frisch (2019). Digital curation through information cartography: A commentary on oral history in the digital age from a content management point of view. *The oral history review 40*(1), 135–153.

Lave, J. and E. Wenger (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Razak, F. H. A. (2008, 2). *Single Person Study: Methodological Issues*. Ph. D. thesis, Computing Department, Lancaster University, Lancaster, UK.

Robertson, B. (2021). Lace version 0.5 new features (draft of 2019-12-21). Technical report, Mount Allison University. `https://tinyurl.com/5fjb6jmw`.

Sharpless, R. (2008). The history of oral history. In R. S. Thomas L Charlton, Lois E Myers (Ed.), *Thinking about Oral History. Theories and Applications, red*, pp. 7–32. Altamira Press.

Social Care Institute for Excellence (2015, October). Co-production in social care: What it is and how to do it. `https://tinyurl.com/5c9et5wz`.

Studdert, D. and V. Walkerdine (2016). Being in community: Re-visioning sociology. *The Sociological Review 64*(4), 613–621.

Taylor, N. and K. Cheverst (2009, December). Social interaction around a rural community photo display. *International Journal of Human-Computer Studies 67*(12), 1037–1047.

The Oral History Society (2022). The history of oral history. `https://www.ohs.org.uk/about-2/`. Accessed: 2022-02-07.

Vayanou, M., A. Katifori, A. Chrysanthi, and A. Antoniou (2020). Cultural heritage and social experiences in the times of covid 19. In *AVI² CH@ AVI*.

Walkerdine, V. and D. Studdert (2012). Concepts and meanings of community in the social sciences.

Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

Willmott, P. (1986). *Social networks, informal care and public policy*. Research Report 655. Policy Studies Institute.

Wilson, J. M. (1882). *The Gazetteer of Scotland*. Edinburgh: W. & A.K. Johnston. Digital version 2020. `https://alandix.com/gzs1882/`.

Zembrzycki, S. (2013). Bringing stories to life: using new media to disseminate and critically engage with oral history interviews. *Oral History 41*(1), 98–107.