

Decoding and Encoding Welsh Manuscript Culture: Scribes, Scripts and TEI

Huw Jones <hej23_at_cam_dot_ac_dot_uk>, Cambridge University Library  <https://orcid.org/0000-0002-8533-9083>
Yasmin Faghihi <yf227_at_cam_dot_ac_dot_uk>, Cambridge University Library  <https://orcid.org/0000-0001-5556-168X>
David Parsons <d_dot_parsons_at_cymru_dot_ac_dot_uk>, University of Wales
Jason Evans <jason_dot_evans_at_llyfrgell_dot_cymru>, The National Library of Wales  <https://orcid.org/0000-0002-8279-3627>
Gruffudd Antur <Gruffudd_dot_Antur_at_wales_dot_ac_dot_uk>, the University of Wales  <https://orcid.org/0009-0007-5383-9028>

DOI: <https://doi.org/10.63744/df9p2wshavdb>

Abstract

Daniel Huws's *Repertory of Welsh Manuscripts and Scribes c.800–c.1800* was published in 2022 to widespread acclaim – a “giant of scholarship” [Russell 2023, 151] which “... handed the world the keys to unlock the great riches of a thousand years of Welsh manuscript culture” [Guy 2024, 95]. Containing descriptions of c.3,300 manuscripts from sixty-eight repositories, along with detailed indexes and plates illustrating scribal hands, the Repertory is “... the kind of work upon which all others build” [Guy 2024, 89]. In his introduction to the Repertory, Huws immediately points to the future, writing that it “... belongs to that class of publication which can only reach maturity in a second edition” [Huws 2022, xxxiii]. This case study describes work towards the production of that edition, not as an expanded or revised set of printed volumes but as a dataset. First, we set the Repertory in the context of catalogues of Welsh manuscripts, and in the wider context of digital resources derived from manuscript descriptions. Then we describe the process of converting the Repertory into a dataset, and how this process, while technical in nature, served as a close reading of its scope and purpose. We offer some practical guidance on the conversion of print volumes to datasets and outline some initial research findings. Finally, we suggest some possible next steps for the Repertory, covering issues such as access and sustainability, integration with other resources and minimal computing approaches to analysing and surfacing the data.

1. Introduction

A *Repertory of Welsh Manuscripts and Scribes c.800–c.1800*, by Daniel Huws, was published in 2022 to widespread acclaim – a “giant of scholarship” [Russell 2023, 151] which “... handed the world the keys to unlock the great riches of a thousand years of Welsh manuscript culture” [Guy 2024, 95]. According to a review in the journal *Manuscript Studies*, “Good reference works compile information in useful ways; great reference works create new pathways for future scholars. In this regard, the Repertory must rank among the best” [Burek 2023, 462]. Containing descriptions of c.3,300 manuscripts from sixty-eight repositories, along with detailed indexes and plates illustrating scribal hands, it is “... the kind of work upon which all others build” [Guy 2024, 89].

In his introduction to the *Repertory*, Huws immediately points to the future:

The *Repertory* belongs to that class of publication which can only reach maturity in a second edition. It bristles with names and numbers: there are bound to be errors, great and small. There will also be errors of judgement with regard to identification of hands, to dating, and other matters.

The work is of a kind that ought to invite additions and improvements: most importantly, in putting names to some of the hundreds of hands which, doubtlessly, with patience and perseverance and luck could be identified from archival sources

[Huws 2022, xxxii].

This case study describes work towards a “second edition” of the *Repertory*, not as an expanded or revised set of printed volumes, nor (in the first instance) as a website, but as a dataset – that is to say a structured and modelled version of the contents of the *Repertory* in digital form. The practical motivations for such work are threefold – the first being, indeed, to publish an online, openly-available version of the data contained in the *Repertory*, the second to take advantage of the affordances of the digital (search, analysis at scale) to provide new views on the data, and the third to link the *Repertory* to the many existing digital resources relating to Welsh manuscripts – from editions and images to biographies, bibliographies, catalogues and dictionaries. However, the work of conversion from printed volumes to digital files also involved a different kind of process – a close reading of the *Repertory*, its structures and connections, its priorities and motivations. The output of this project is a resource which is practically useful in researching Welsh manuscripts, and a dataset which represents a specific model of a particular view of Welsh manuscript culture embedded in its own historical context. This dual nature of the *Repertory* presents challenges which are relevant to digital manuscript catalogues specifically and are also generalisable to many processes of data modelling in a humanities context.

In his review of the *Repertory*, Ben Guy emphasises that it is “[s]o much more than merely a catalogue, it is a powerful tool for both research and teaching” [Guy 2024, 88]. One of the main aims of the project was to allow scholars to make full use of this potential, not just by increasing access to the *Repertory*’s contents but also by facilitating new views on the data. However, it was also crucial that the dataset retained the original vision and scope of the printed volumes. This meant that the automated work of structuring, dissection, identification and linking by which the dataset was created was less a process of data modelling *per se*, and more an exercise in understanding and drawing out the implicit models and structures already present in the *Repertory*.

Some of these (the linking of a name in an index to a manuscript) were relatively straightforward to model automatically – others (the particular role of a person in relation to a manuscript) less so. Inevitably, we relied enormously on collaboration between team members to make sense of the *Repertory* and its contents. The creation of the dataset was a hybrid of automated processes and traditional humanities research – with a conversation between technical expertise and subject-specific knowledge at its heart. The *Repertory* does, indeed “bristle with names and numbers,” but there are remarkably few errors in its structures and linkages. The real work lies in understanding the volumes and ensuring, as far as possible, that the affordances of the dataset match the vision implicit in the indexes and entries. This engagement between technical processes and manuscript expertise was the core of the project, and its most fulfilling and enjoyable aspect.

This case study begins by putting the *Repertory* into context, both among catalogues of Welsh manuscripts and more broadly among datasets derived from manuscript descriptions. We also touch on previous research on the construction of datasets from catalogue data, including our own work on quantitative approaches to collections of Islamicate manuscripts. We walk through the methodology used in the conversion process, highlighting areas which presented particular challenges such as the parsing of dates and the automated linking of index entries to manuscript descriptions. This section concludes with a brief description of the concrete outputs of the project and an outline of some initial research findings.

In our methodological conclusion, we offer practical guidance on the conversion of print volumes to datasets and reflect on the close collaboration between manuscript research and technical expertise which drove the project. In the final section, we outline possible next steps for the *Repertory* as a series of technical interventions which aim to facilitate greater (and more diverse) access to and engagement with the dataset – particularly in the context of Welsh as a minority culture and language. Drawing on the “full stack” framework proposed by Alan Liu [Liu 2020], we cover a number of core issues which are likely to face the *Repertory* in the next stages of its development including access and

sustainability, integration with other resources, and minimal computing approaches to analysing and surfacing the data.

2. RELATED RESEARCH

2.1 The Repertory and its context

Catalogues give manuscripts a voice, but the language in which they speak varies according to the interests and priorities of the scholars who catalogue them [Formigatti 2017, 29].

8

Although by no means the first catalogue of Welsh manuscripts, the *Repertory* possesses a unique vision, a claim to completeness in terms of that vision, and an up-to-date and authoritative voice which sets it apart from what came before, and from most comparable work on other manuscript cultures. Huws describes the scope of his work as follows:

9

The aim of the *Repertory* is to include all manuscripts up to 1800 that belong to the Welsh tradition. That is, Welsh literature and scholarship, British history as seen by the Welsh, genealogy, Welsh law, science and medicine in so far as there was a continuum of learning in Welsh from medieval to modern times. The tradition includes manuscripts relating to these fields, and written from a Welsh standpoint, that are in Latin or English

[Huws 2022, xxxi].

The work's main predecessor – the only previous union catalogue of Welsh manuscripts from different repositories – was the *Report on Manuscripts in the Welsh Language*, published by J. Gwenogvryn Evans between 1898 and 1910. Gwenogvryn's work was painstaking and his lists of the manuscripts' contents are very full, often fuller than the summaries presented in the *Repertory*, a point taken up below. Inevitably, however, his catalogue shows its age in outdated identifications of scribal hands and assessments of date and provenance. Moreover, it deals with only around a quarter of the material included by Huws. Other earlier catalogues relate to single repositories (particularly the National Library of Wales and the British Library) and share the problem of including obsolete and inaccurate information.

10

Huws's achievement has been to assemble a much more complete corpus, and to provide a much more accurate level of information than has previously been available. As former Keeper of Manuscripts in the National Library of Wales, he could bring decades of his own expertise to the project, besides being able to draw on an up-to-date familiarity with a century of modern published research in the field. Huws particularly concentrates detailed attention in the *Repertory* on the identification of the scribes and owners of the manuscripts, adding extensively to existing understanding, and opening up countless research pathways into networks of knowledge and influence. As the reviews quoted above have emphasised, the *Repertory* makes Welsh manuscript culture – the written culture of a nation across a millennium – available for study in new and exciting ways.

11

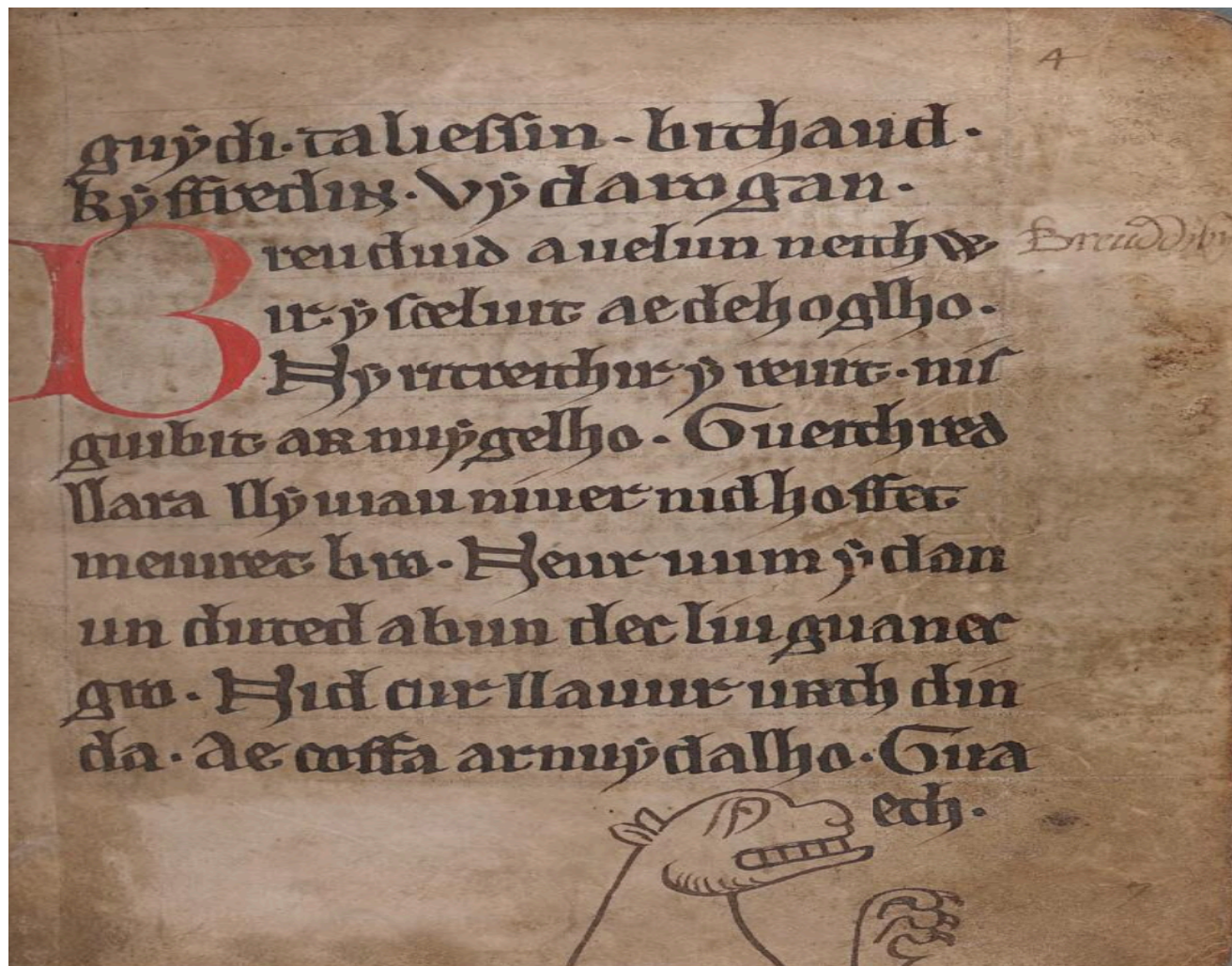


Figure 1. NLW, MS Peniarth 1, *Llyfr Du Caerfyrddin*, 4r.

All that said, there is no suggestion that the *Repertory* should be the final word, as the author acknowledges in his comments on a second edition, quoted above. New discoveries are likely to be made in repositories around the world (and, indeed, are being made: at least 30 additions have come to light in the wake of the *Repertory*'s publication). Moreover, the chronological and thematic limitations Huws imposed on himself leave “ragged edges” [Huws 2022, xxxi], in excluding material that by other definitions of Welsh written culture might be included. Paul Russell’s review of the work [Russell 2023], whilst lauding the “extraordinary achievement,” lays some emphasis on this aspect: he observes that there might be room for much more archival material, cartularies and, more specifically, “medieval manuscripts in Wales in other languages, those in Latin, French, Middle English and so on,” a point that had been conceded by Huws, lamenting the exclusion of “the fifty or so medieval manuscripts of known Welsh provenance which do not contain any Welsh text” [Huws 2022, xxxi].

12

2.2 Related Resources

If there is scope in the future to add further manuscripts – whether new finds belonging to the ‘Welsh Tradition’ or works admitted under extended criteria – there is also considerable potential to link the contents of the *Repertory* to other resources online. Images of the manuscripts would be one particularly attractive addition. The third volume of the published work contains a thousand plates illustrative of different scribal hands. While it may be possible to reproduce these online (though there are copyright complications which go beyond digitising the text), there would still be considerable and evident value in linking to fully digitised manuscripts where they are available (roughly 400 have been digitised from the collections of the National Library alone). In turn, it might be hoped that the habitual linking of *Repertory* entries and manuscript images will strengthen arguments for the digitisation of more manuscripts.

13

Recent projects in Welsh studies have included the full transcription of some medieval manuscripts (<https://www.rhyddiaithganoloesol.caerdydd.ac.uk/en/>), as well as critical editions of a number of literary texts, all of which include transcriptions from numerous manuscripts (e.g. <https://www.gutorglyn.net/>, <https://dafyddapgwilym.net/>, <https://saints.wales/>). Linking from the entries on manuscripts in the *Repertory* to the relevant sections of these sites will begin to build a rich contextual superstructure. Further digital editions of texts are currently in development, and it might be hoped that reciprocal linking to the *Repertory* online will become a usual expectation of such work.

14

It was noted above that since Huws concentrates particular attention on the scribes and early owners of manuscripts, he is less concerned with a forensically detailed breakdown of the contents. Such a breakdown is available on the online catalogues of some repositories, generally derived from the old print works, and it would be possible to link to these. It would also be possible to create digital versions of the old catalogues, especially Gwenogvryn Evans's work, and to build links between entries. In these cases, there are questions to face about data provenance and hierarchies of authority: serving up old material with opinions on dating and attribution which may conflict with, and should be superseded by, the *Repertory*, would need to be negotiated carefully.

15

The National Library of Wales currently holds one other resource, which sits somewhat between those mentioned above, and could be usefully integrated with an online *Repertory*. Maldwyn is a detailed index to Welsh poetry in the manuscripts, a dataset of some 50,000 records, with each item ascribed to its manuscript. Practically all of these will be in the *Repertory* collection, and the detail could be usefully linked – though again there are questions surrounding, for example, the accuracy of attributions to named poets. There are obvious links to be made with other major resources in a Welsh context – many of the people referenced in the *Repertory* will be recorded in the Bywgraffiadur Cymreig, and many of the places will be indexed in Coflein. Experimental work to link these datasets is described in section 4.6.3.

16

Outside of a Welsh context, there are obvious links to be made with related catalogues of manuscript material. Our use of a shared data model makes the data immediately compatible with the TEI catalogues maintained by the Bodleian Libraries, including Medieval Manuscripts in Oxford Libraries and FIHRIST, and with the digitised collections of Cambridge, Manchester and Lancaster Universities. A certain amount of cross-walking will open up the potential for cross-analysis with the collections of the British Library and others. Here, resolving the identity of people, places and works is likely to be more of a challenge than working with different data formats.

17

2.3 Manuscript Catalogues as datasets

Paul Russell's review expands Huws's own assertion that the *Repertory* is the starting point for further work – that it "... should be seen not as the completion of a project but as the beginning of a larger one" [Russell 2023, 155]. It is this which presents the most complex and interesting challenge to the creation, curation and maintenance of the *Repertory* as a dataset. Datasets change in ways which are less obvious and harder to trace than editions of printed works. If additions and alterations to the dataset are both desirable and inevitable, at what point does it cease to be Daniel Huws's *Repertory of Welsh Manuscripts and Scribes* and start to be something else?

18

This question has real implications for users of the *Repertory*. A citation to a printed edition remains relatively stable. But what happens if an article refers to an entry in a dataset which then changes – for instance, the redating of a manuscript or the identification of a previously unknown scribe? Ben Guy points to several instances where "there have been some slight changes of opinion about certain manuscripts by comparison with earlier lists" [Guy 2024, 92]. If opinion changes, what happens to the "original" information? Should it be overwritten, or maintained as also being of interest to scholars? And how can its status be recorded? In the analogue world researchers know exactly which edition of which book they are using, but the sources and provenance of information can be less clear in the digital domain:

19

Crucially, the historically specific labours and practices of catalogue production are all too easily obscured by the presentation of the catalogue as an always-already present unifying entity, and are further obscured when collections are federated for access, into datasets, or as machine readable endpoints

One approach is to see the digital resource as primarily an edition of the *catalogue* rather than directly as a source of information about the items described. This was the method adopted by *Enlightenment Architectures* [Sloan and Nyhan 2021], which created digital editions of five of Hans Sloane's catalogues to a) analyse the language that was used to describe items in his collections, and b) attempt to reconstruct his original collections from the contents of the catalogue descriptions. There are many parallels between this project and ours, including the use of TEI as a data standard, a focus on close reading of the catalogues as the basis of the creation of the dataset, and the iterative nature of the data analysis. However, the most important similarity is that both projects have collaboration at the centre – in their case between “[r]esearchers – curators, librarians and university academics – with contiguous knowledge specialisms in the histories of the book, of natural history, of antiquities and ethnography, and of collections, as well as historians of science and information science” [Sloan and Nyhan 2021, 199]. This kind of interdisciplinary engagement is crucial to good work in the digital humanities.

20

The Sloane catalogues were manually transcribed into TEI and semantically encoded in response to the project's research questions – focusing on people, places, bibliographic references and descriptive terms. The project differed from ours in that the source data (Sloane's manuscript catalogues) were primarily of historical interest as documents in themselves – so the model only had to represent the sources in their present state. This meant that while the encoding might change and develop over time to draw out new aspects of the documents, there was no need for the model to include a sense that the documents *themselves* might be subject to change.

21

An alternative is to treat the catalogue primarily as a source of data *about* the manuscripts – and one that will be edited and added to as our understanding of those manuscripts develops. This model implies an acceptance that the dataset will change, and requires strategies for recording the nature and provenance of those changes, as well as a means of retaining and retrieving previous versions of the data. This is the approach adopted by the FIHRIST Union Catalogue of Islamicate manuscripts, and by extension the TEI datasets which share the same model – including the nine TEI-based manuscript catalogues supported by the Bodleian Libraries, and the datasets which underlie the digital library platforms of Cambridge, Manchester and Lancaster universities. These are all “live” datasets (they are being edited and added to on an ongoing basis) not just in terms of encoding new sources, but also in terms of changing existing descriptions – either as a result of new research or by the integration of information from other resources. These datasets become an agglomeration of material derived from different sources, different activities over time, and so require a different approach to data modelling and data provenance.

22

We have written at some length about FIHRIST as a dataset, specifically as a basis for quantitative approaches to codicological analysis [Faghihi and Jones 2024], but it is worth reemphasising a few key points. Like the *Repertory*, FIHRIST is a union catalogue covering the collections of multiple institutions. Unlike the *Repertory* it is not the product of a single authorial/editorial effort, but an aggregation of the results of multiple cataloguing processes over time, including the conversion of legacy catalogues and original cataloguing from manuscripts themselves. The dataset continues to be added to and edited in a collaborative workflow managed through GitHub. In this context, the provenance of the information in the dataset takes on extra importance, especially when analysing the data at scale.

23

The *Repertory* sits somewhere between the two approaches. On the one hand it is a work complete in itself with its own context and purposes – on the other its creator explicitly sees it as a work “... of a kind that ought to invite additions and improvements” (though these are initially envisaged as an extension of the original purpose of the *Repertory* – “... in putting names to some of the hundreds of hands which, doubtlessly, with patience and perseverance and luck could be identified from archival sources” – rather than using the data to answer different kinds of research questions). The idea that the *Repertory*, in its dataset form, would immediately begin to change and develop, led us to follow the FIHRIST data model – a dataset of manuscript descriptions rather than a digital edition of a manuscript catalogue.

24

3. METHODOLOGY AND OUTPUTS

The *Repertory* is published in three volumes – the first containing summary descriptions of c.3,300 manuscripts, the second biographies of scribes and indexes to place names, personal names, texts and subjects, and the third images providing examples of scribal hands. This project concentrated on the first two volumes – the manuscript descriptions

25

and the indexes, in the anticipation that the availability of images of manuscripts included in the *Repertory* over time might allow for a more comprehensive approach to the illustration (and further identification) of scribal hands in the future.

3.1 General Considerations

There were several features of the project and the source material which had a fundamental effect on how we approached the conversion from print to digital.

The first, and most important, was that the conversion was a one-off process. As the source was fixed and not liable to change, there would be no need to run the process multiple times. Once the print version of the *Repertory* had been satisfactorily converted into digital format, this would be a new source of truth. This meant that manual intervention, both in correcting the source (where minor errors of formatting or punctuation affected the automated procedures) and in dealing with extreme edge cases (very small numbers of cases which it was more efficient to fix manually than to write code), could be part of the process. This saved a significant amount of time and extra coding – but put huge emphasis on checking the outputs of our work. If we were only going to run the conversion once, we had to get it right.

The second was the comparatively small size of the dataset – c.3,300 manuscript records and five indexes. This meant that a certain amount of manual intervention was possible for tasks which were difficult or impossible to automate. It also opened the way to semi-automated approaches – a lightweight combination of coding and manual correction which reduced the amount of code we needed to run (removing the need to cater for a long tail of edge cases) and the number of iterations in the process. This approach proved very useful in the parsing of important aspects of the data such as dates and classmarks. In larger datasets such as FIHRIST (15,400 records) and Cambridge Digital Library (49,524 records) manual interventions (even for edge cases) would be almost impossible because of the amount of work involved.

The third is that unlike conversions of historical printed catalogues, we were able to work from digital files. The usual starting point for conversion is either scanning of the printed volume(s) followed by OCR – which usually involves a considerable amount of manual checking and conversion – or manual transcription, which is extremely time-consuming. We worked from the InDesign files used to publish the printed volumes, which saved a huge amount of time and effort at the start of the project and also gave us a reliable data source. InDesign supports several export formats, and after some experimentation we settled on HTML as an XML-like standard which, while directed towards display rather than semantic markup, retained machine-readable versions of all text formatting.

One important decision we made at an early stage was that each manuscript description should have its own separate TEI record (instead of one large file containing all the manuscript descriptions). This follows the model adopted by FIHRIST and is in part influenced by considerations around collaborative workflows. While GitHub has mechanisms for resolving instances in which one file has been edited by more than one person, it is generally advisable to avoid situations where multiple changes to a single file must be manually resolved. A large file containing all manuscript descriptions would require the constant merging and resolution of changes made by different editors. By comparison, editing of index entries in authority lists (which were kept as single files for each index rather than separate files for each entry) is comparatively uncommon and can be treated as a discrete editorial procedure. Again, these decisions envisage a future where the curation and expansion of the *Repertory* is an ongoing and collaborative undertaking with editorial oversight.

We also decided to retain the original entries for manuscripts and index entries as part of the final dataset, in addition to the more specific and granular fields which were created because of the data parsing. This was done partly to ameliorate the risks of the one-off conversion process (we can always go back to the source data in the event of mistakes or for further data extraction), and partly to allow us to create an online publication which, while having the extra search and faceting capabilities facilitated by the enriched data, would also allow us to display Daniel Huws's descriptions as they were originally published in the print volumes.

In terms of technologies, all coding was done in Python and XSLT. For the Python code we made heavy use of the

Beautiful Soup library for parsing and writing HTML and XML. We also looked at the LXML library, which has better support for XPath, but is a little less user-friendly in dealing with namespaces and some other XML features. We tended to use Python in the initial stages of the conversion, taking advantage of its text processing capabilities to break the files up into their constituent entities and to parse specific aspects of the data (dates, classmarks, names etc.). Then, once we had established the general structures, we used XSLT to iteratively correct and analyse further elements of the descriptive records and indexes. While it would have been possible to do everything in either Python or XSLT, this mixed approach allowed us to play to the strengths of each language. In general, Python allows more flexibility and has a wider range of functions for text manipulation, while XSLT is more straightforward and easier to test when changing large quantities of files. We ran the XSLT using the Oxygen XML Editor, but it would also have been possible to use open-source alternatives, or to run conversions from the command line. We made significant use of two open tools provided by the TEI community itself – the TEIGarage conversion tool for the initial conversion from HTML to basic TEI, and the ROMA ODD editor for the creation and editing of the schema.

We have talked about the importance of a collaborative close reading of the *Repertory* as the basis for both the data model and the conversion. The first and most fundamental stage of the project was to spend a lot of time with the *Repertory*, getting to know its structures and mappings, its quirks and idiosyncrasies. This process involved an ongoing conversation between technical expertise and subject knowledge which continued through the life of the project. Also important was Daniel Huws's detailed introduction, where complex aspects of the data such as formats for dates were clearly explained, and Paul Russell's review of the *Repertory* [Russell 2023], which pointed out some crucial features such as the treatment of unknown scribes whose hands only appear in one manuscript. This close reading, accompanied by detailed discussion, allowed us to make important decisions about the data model at an early stage – including the combination of the scribes and personal names indexes into one file, the splitting of the titles and subject index into two indexes, and the retention of the full original entries for both manuscript records and index entries as part of the dataset.

33

3.2 Workflow

In terms of general workflow, we began by converting the manuscript records, and then went through each of the indexes in turn. For each class of information (manuscripts, persons, places, repositories, subjects, works), the overall workflow followed the same pattern:

34

- Export from InDesign to HTML
- Conversion in TEIGarage of HTML to basic TEI encoded for display
- Processing of the TEI into identifiable segments using layout (spacing, columns) and formatting (capitals, superscript, italic, bold, etc.)
- Establishing data categories for each segment using a combination of order of occurrence, formatting and textual content
- Where possible, processing human readable values into machine readable formats (e.g. dates into ISO format)
- Assignment of unique identifiers to each entity in the dataset (manuscripts, people, places, repositories, collections, subjects, works)
- Mapping between indexes and manuscript entries
- Output as a semantically-encoded TEI dataset

In the following sections we draw out some of the more interesting and challenging aspects of the conversion workflow.

35

3.3 Structuring and segmenting

Having converted the HTML outputs of InDesign into basic TEI, the first task was the identification of discrete entities (a manuscript, a person, a place), followed by the identification of structures internal to each entity (a date, a title, a name, a classmark). Broadly speaking we employed two methods for this process.

36

The first was to use the formatting and ordering inherent in the files to break the text into segments. Entries tended (with

37

some exceptions) to start and end in the same way. Within entries, some structures could be identified by position (the first line of a manuscript entry was a header consisting of the classmark, followed by a line with the title and date, the first line of a person entry was the standard form of the name followed by dates) and some by formatting (the header of each manuscript record was in bold, the physical descriptions of manuscripts were blocks of text in small capitals, some aspects of dates always appeared in superscript).

Llanstephan 8 (15B)

Brut y Tywysogion

s.xvii¹

Brut y Tywysogion, a text of the 'Peniarth 20 version' made good by a text related to Pen. 19 [see Jones, *BT (P20 tr.)*, p.xxiv; Jones, *BT (RBH)*, p.xxxiv]. All calligraphically written in the hand of **George William Griffith**, employing his secretary, italic, roman and legal scripts (his initials on fol.131^v). Extensive side-notes in English, including much quotation from a variety of other sources, notably David Powel, *The Historie of Cambria* (1584). A continuation in Welsh up to AD 1544. The text was copied in 1717 by James *Davies (Iaco ab Dewi), see Card. 2.135.

On fol.5^v, the signature of Thomas Lewes, s.xvii/xviii. Given to Moses *Williams by W. Lewis (see his list of texts in Llst. 57). Inside the cover, a note in the hand of William *Jones.

4°. 147 fols. Gold-tooled light calf, s.xviii¹, as Llst. 1–4; lettered 'PRYD TYWYSOG CYMRU'.

Figure 2. Sample manuscript description from printed volume. Note use of spacing, order, bold, italics and small capitals to indicate different parts of the entry.

The second was to use the textual content of the segments to establish their semantic value. In the physical description, extent statements tended to include some abbreviation indicating the number of folios, binding statements tended to include one or more of a number of common words.

38

```
#these usually indicate extent
elif re.search("\d [Ff]ol[\. ]|\d fol[s]*$\d [Ff]ols[\. ]|membrane|sheet", physdesc_part):
    extent.append(physdesc_part)
    previous="extent"
```

Figure 3. Example regular expression to find text blocks indicating extent.

In practice, a combination of rules regarding formatting, position and textual content were often combined in the code to assign meaningful values to segments. The comparatively long (134 lines of code) Python subroutine which separates the physical description into distinct fields for material, extent, collation, foliation, binding and notes is a good example of

39

a combined approach to text parsing.

The data parsing was a collaborative and iterative process which relied on conversations amongst team members and on guidance found in the *Repertory* itself. For instance, the parsing of dates from the formats used by the *Repertory* into machine-readable single dates or ranges (one of the most complex sections of the code, running to some 483 lines of Python) was in essence an encoded version of the detailed explanation given in Huws' introduction. In a similar vein, the parsing of classmarks, repositories, abbreviations and many other aspects of the data relied on the comprehensive guidance in the pages of the *Repertory* along with consultation with subject experts.

```
#a range of centuries with or without qualifiers, including quarter century qualifiers
#e.g. s.xvii-s.xviii, s.xvii2-s.xviiiimed, s.xviii1/4-s.xviiiimed

elif re.match("^s\[xvi\|]+\[dmed\|*\[\/\|]\[xvi\|]+\[dmed\|]*$", date):
```

Figure 4. . Example regular expression from date parsing subroutine.

A challenge which came up several times in the parsing process was the different format used for some aspects of the data in different parts of the *Repertory*. Classmarks, for instance, are recorded slightly differently in the manuscript records and the indexes. There are subtle variations between the date formats used for people and those used for manuscripts. The forms of names used in the Scribes and Personal Names indexes are also very slightly different. These variations, which are clearly explained in the introductions to volumes, are trivial or even invisible to users of the printed volumes – it is always clear to a human reader what is being referred to. For a computer, however, even small variations need to be identified and analysed before parsing and mapping between entities can be achieved.

```
<persName type="display">Lloyd, Dafydd (Dafydd Llwyd Bach), Mer., 18<hi
rend="superscript">2</hi> cent.</persName>
<persName subtype="scribe_index" type="variant">LLOYD, Dafydd (Dafydd Llwyd Bach)
(s.xviii<hi rend="superscript">2</hi>).</persName>
```

Figure 5. Encoded entries from personal name (top) and scribe (bottom) indexes – note the inclusion of place (Mer.) in first entry and the difference in date formatting.

3.4 Mapping and identification

A fundamental part of the conversion process was the assignment of unique identifiers to each entity in the dataset – manuscripts, people, places, repositories, subjects and works. These identifiers were used to map the relationships between entities – between a person and a manuscript, a manuscript and a repository, a manuscript and a place, a work and a person, etc. Each identifier has a prefix indicating what kind of entity it is – e.g. manuscript_, person_, followed by a running number, so manuscript_1245, person_626. Identifiers are kept purposefully without meaning (e.g. they don't have to follow the alphabetical sorting of the indexes or entries) to allow for expansion of the dataset without reassignment of identifiers – though in a shared editing environment this does require a workflow to avoid duplication of identifiers. The assignment of editorial work on the indexes to a single team member is one way of addressing this problem.

One of the challenges we faced was how to represent the hierarchical nature of the places, works and subject indexes. Here, the main entry is indexed under the running number (e.g. "Bible, Welsh" is subject_24) with subcategories indexed under the main entry, followed by a running number within that entry (e.g. "Bible, Welsh – bibliography" is subject_24-2). This allowed us to follow the model of having a unique identifier for each entity in the dataset, while at the same time retaining a sense of the entity's place in the hierarchy in both machine and human-readable formats.

```

<item xml:id="subject_28">
  <term type="display">bindings (<hi rend="italic">binding fragments in general, not
    indexed</hi></term>
  <note type="category" xml:id="subject_28-1">
    <term type="display">medieval (intact or in part)</term>
  </note>
  <note type="category" xml:id="subject_28-2">
    <term type="display">16 cent. (intact or in part)</term>
  </note>
  <note type="category" xml:id="subject_28-3">
    <term type="display">17 cent.</term>
  </note>
</item>

```

Figure 6. Hierarchical subject index entry with xml:id attributes indicating that sub-categories belong to overall heading.

For mapping between entities we followed the explicit references between manuscript entries and indexes present in the printed volumes – so if a person is indexed against a manuscript, a mapping is made, if a place is indexed against a manuscript, a mapping is made, etc. Having resolved some problems with the different ways in which classmarks and repositories are formatted in the manuscript entries and indexes, this was a comparatively straightforward process. However, there are many relationships implicit in the indexes and the entries which are not explicitly mapped – both internal to indexes (people are cross-referenced in each other's entries in the persons index) and across indexes (relationships between people and places are present but not explicitly mapped). The explicit mapping of these implicit relationships would greatly enrich the research value of the dataset, but would be difficult to automate other than through AI-based natural language processing techniques – which are prone to error even for straightforward texts (confusion between personal and place names being one prominent example). Semi-automated solutions to this problem may form part of future work on the dataset.

44

3.5 Semi-automated and manual processes

As outlined above, the comparatively small size of the dataset enabled us to incorporate semi-automated or entirely manual processes which would have been impossible with a larger corpus. Manual interventions raise interesting questions regarding conversion work. Entirely manual approaches are time consuming even for smaller datasets and there is a balance between the resources spent on manual adjustments to the data and the benefit such adjustments bring to users of the dataset. On the other hand, there is an efficiency in the use of semi-automated approaches for data where there are a lot of individual edge cases.

45

The clearest example of a semi-automated approach was date parsing. Following the detailed guidance in the introduction to the *Repertory*, it was possible to automatically parse the vast majority of dates. In essence, date parsing was an exercise in pattern recognition of combinations of the various date formats used in the volumes. The subroutine which did the date parsing was one of the more complex aspects of the coding – 483 lines of Python identifying seventeen main date combinations, each with multiple subclauses. However, even this level of coverage left a long-ish tail of edge cases (c.30 date formats), each of which only occurred once in the dataset. One option would have been to add a further 30 clauses to the subroutine, but this would have been much more time consuming than manually fixing the edge cases.

46

This approach (which saved us a lot of time across various processing tasks – parsing of the various classmark and repository formats being the other prominent example) would have been impossible had there been a requirement to run the conversion multiple times – in which case we would have needed to remember to manually correct the edge cases every time we ran the conversion. The fact that this was a one-off conversion, coupled with the size of the dataset, allowed us to build in a small amount of manual correction to the automated processes to save time and prevent the needless proliferation of code.

47

Some important tasks were impossible to automate and had to be addressed manually. One aspect of the persons index which is not explicitly brought out in the data (though it is implicit to readers of the entries) is gender. We decided to add gender as a machine-readable value partly because it was felt that it could facilitate interesting new analyses of the corpus (were women involved in the writing/collection of particular kinds of texts, did the involvement of women in the production or collection of Welsh manuscripts change over time), and partly as a test case to estimate the time necessary to make manual adjustments to the dataset. A brief overview of the persons index indicated, as one might expect, that the vast majority of the entries referred to men. As a starting point, a default attribute value `sex="M"` was added to every person element in the index. Then we looked through the index and changed the value to `"F"` for entries relating to women. There are 8,349 entries in the persons index (7,871 male, 478 female) – and adjusting the values took about two days for a single team member with a familiarity of Welsh personal names and of XML – a significant time commitment, but not one which would discount further manual adjustments or additions of this type.

48

A much larger manual task was the splitting of the “Subjects and Titles” index into separate indexes for subjects and works. This was a lengthier process, taking one team member just over four days. It involved the manual parsing of the 1,380 entries in the Subject and Title index and their allocation to separate files. While this was a significant time investment for the project, it was felt to be sufficiently valuable not just in terms of data compatibility but also in allowing a more granular approach to the dataset and to the potential research questions which could be asked of it (did certain scribes write on particular subjects, were certain texts copied in particular periods). It also allowed us to make explicit mappings between the persons and works indexes to indicate authors, editors and translators of texts (as distinct from scribes, owners, annotators of particular copies of texts).

49

A manual process which is still in progress is the formal encoding of the types of relationships between people, places and manuscripts. While the indexes indicate connections between people, places and manuscripts, the nature of those relationships are not explicitly mapped – and it is very difficult (even with natural language processing tools) to automatically establish what type of relationship exists between the entries – e.g. whether a person is the author, scribe, translator or owner of a manuscript, whether a place is where the manuscript was copied, what the manuscript is about, or where the manuscript is known to have been, etc. This kind of work requires subject expertise in the Welsh manuscript tradition, and it was felt that it would be best accomplished by researchers with a knowledge of the field. We held a series of short, intense training sessions for team members to familiarise them with the tools, standards and technologies required to edit the data, and workshopped the construction of controlled vocabularies for relationships between entities (for people we made use of a subset of the Library of Congress MARC relator codes, for places we constructed a project-specific vocabulary). These values were embedded in the schema for ease of editing.

50

While the addition of codes to each of the c.3,300 manuscript records in the corpus is likely to take a considerable amount of time and effort it was considered an invaluable addition to the dataset, without which it would be impossible to usefully interrogate the kinds of complex relationships inherent in the data. The training sessions and workshop discussions also had the advantage of engaging the researchers with the technical aspects of the project and familiarising them both with the data itself and with tools and platforms such as the Oxygen XML Editor and GitHub. Too many digital humanities projects fail due to lack of knowledge and skills transfer from the people who built the dataset to those who will be expected to look after it in the long term.

51

3.6 Outputs

The concrete outputs of the project are hosted and documented on the project GitHub site. Access is currently restricted to project members but will be made openly available with the launch of the dataset towards the end of 2026.

52

3.6.1 Dataset

A TEI-encoded dataset consisting of:

53

- 3,304 manuscript records
- 8,600 name records
- 1,726 place records

- 68 repository records
- 273 subject records
- 1,107 works records

Records were parsed to extract granular data values where possible (e.g. title, date, classmark, material, extent, binding, collation, foliation etc.) while retaining the original entries for completeness and further processing.

54

```
<physDesc>
  <objectDesc>
    <supportDesc material="paper">
      <support>
        <material>4<hi rend="superscript">o</hi>.</material>
      </support>
      <extent>147 fols.</extent>
    </supportDesc>
  </objectDesc>
  <bindingDesc>
    <p>Gold-tooled light calf, s.xviii<hi rend="superscript">1</hi>, as
      Llst. 1-4; lettered 'PRYD TYWYSOG CYMRU'.</p>
  </bindingDesc>
</physDesc>
```

Figure 7. Encoded physical description of NLW Llanstephan 8, showing granular values for material, extent and binding.

Selected values (dates, gender, classmarks) were parsed into machine readable values for search, sorting and analysis.

55

```
<origin>
  <origDate calendar="gregorian" notAfter="1650" notBefore="1600">s.xvii<hi rend="superscript">1</hi>
</origDate>
</origin>
```

Figure 8. Encoded origin section of NLW Llanstephan 8, showing text version of date automatically parsed into ISO values in the notBefore and notAfter attributes.

Each record in the dataset was identified with a unique identifier using the xml:id attribute. Relationships between entities as recorded in indexes were mapped using key attributes.

56

```

<listPerson>
  <person role="scr">
    <persName key="person_800">Davies, James (Iaco ab Dewi;
      1647/8–1722)</persName>
  </person>
  <person role="scr">
    <persName key="person_1991">Griffith, George William, Penybenglog
      (1584–1655)</persName>
  </person>
  <person role="scr">
    <persName key="person_3675">Jones, William, FRS, Llanbabo (<hi rend="italic">c.</hi>1675–1749)</persName>
  </person>
  <person>
    <persName key="person_4025">Lewes, Thomas, 17/18 cent.</persName>
  </person>
  <person>
    <persName key="person_4037">Lewis, W., 17<hi rend="superscript">1</hi>cent.</persName>
  </person>
  <person role="scr">
    <persName key="person_7910">Williams, Moses (1685–1742)</persName>
  </person>
</listPerson>

```

Figure 9. listPerson entry in manuscript description for NLW Llanstephan 8 showing mapping of people to manuscripts using key attributes referencing xml:id attributes in persons index.

3.6.2 Schema

A TEI ODD file and RNG schema were created using the ROMA ODD editing tool and populated with controlled vocabularies to support the editing workflow and facilitate data validation.

57

```

<listPerson>
  <person role="">
    <persName key="">Former owner of the manuscript</persName>
  </person>
  <person role="scr">
    <persName key="">Davies, John, Mallwyd (<hi rend="italic">c.</hi>1567–1644)</persName>
  </person>
  <person role="scr">
    <persName key="">Gryffyth, Jaspar, Guilsfield (<hi rend="italic">c.</hi>1560–1614)</persName>
  </person>
  <person role="scr">
    <persName key="person_2717">Humphreys, Humphrey (1648–1712)</persName>
  </person>
  <person role="scr">
    <persName key="person_3053">Johns, David, vicar of Llanfair Dyffryn Clwyd
      (<hi rend="italic">fl.</hi>1569–89)</persName>
  </person>

```

Figure 10. Use of controlled value list embedded in schema for editing workflows and data validation.

3.6.3 Wikidata/SNARC integration

Partial alignment of authority entities to Wikidata and the National Library of Wales' Name Authority Wikibase was achieved using OpenRefine to semi-manually reconcile personal names and place names against those in Wikidata.

58

8500 rows Schema Issues Preview						
Show as: rows records Show: 5 10 25 50 100 500 1000 rows						
		person - xml:id	person - persName	Name	birth	death
16.	person_16	Annes verch Robert ap John ap Ithell	Annes verch Robert ap John ap Ithell	Search for match		
17.	person_17	Anni Goch, 15/16 cent.	Anni Goch	Search for match		
18.	person_18	Aron ap Gr' ap Aron, 15 cent.	Aron ap Gr' ap Aron	Search for match		
19.	person_19	Arthall	Arthall	Search for match		
20.	person_20	Arthur, king	King Arthur	Search for match		
21.	person_21	Arthur ab Dafydd	Arthur ab Dafydd	Search for match		
22.	person_22	Arthur ap Huw	Arthur ap Huw	Search for match		
23.	person_23	Abel, Sion, 18 cent.	Siôn Abel	Search for match		
24.	person_24	Abraham (Abram), Richard (Dic y Dawns)	Richard Abrah	Search for match		
25.	person_25	Acton family, Salop	Acton family	Search for match		
26.	person_26	Acton, Oliver (	Oliver Acton	Search for match		
27.	person_27	Adams, Richard, Ludlow (	Richard Adams	Search for match		Ludlow
28.	person_28	Adams, Richard, 218 cent.	Richard Adams	Search for match		
29.	person_29	Adams, William (	William Adams	Search for match		
30.	person_30	Adamson	Adamson	Search for match		
31.	person_31	Addams-Williams family I lannibibw Castle	Addams-Williams family	Search for match		I lannibibw Castle

Figure 11. Use of OpenRefine to aid reconciliation to Wikidata as a controlled vocabulary.

Once reconciled, the unique identifiers from the TEI were added as external identifiers to the relevant data items in SNARC. This made it possible to access additional biographical and geographical data, as well as links to other sources of authority such as VIAF and Library of Congress.

59

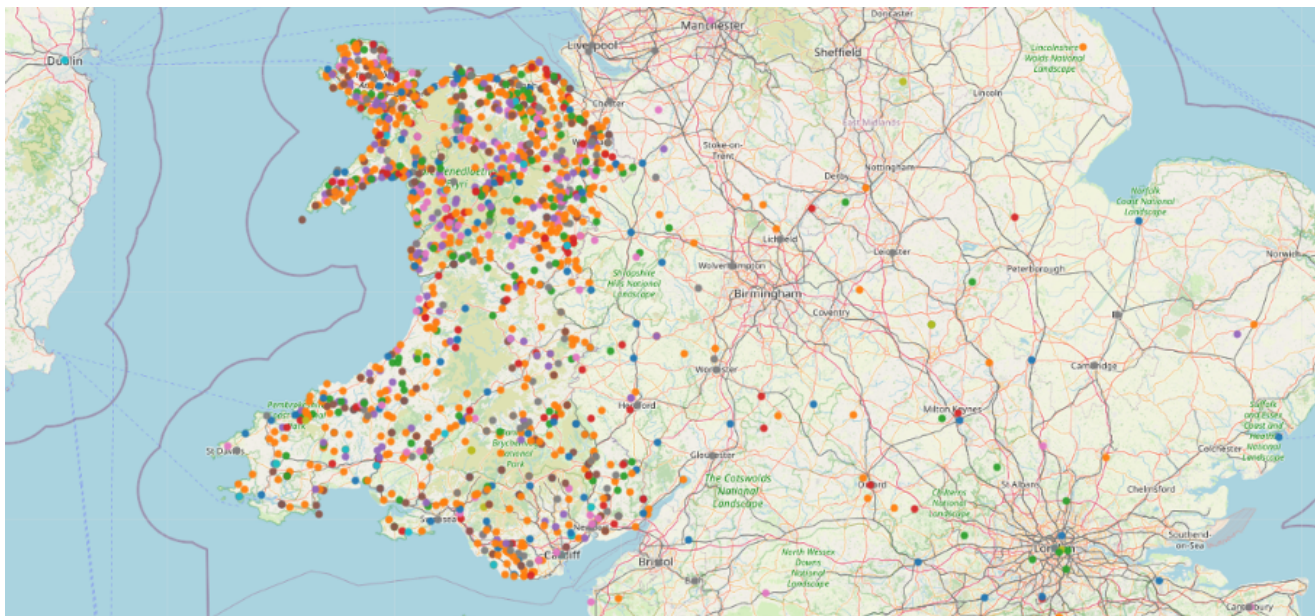


Figure 12. Map of place names mentioned in the *Repertory*, colour coded by type.

This process also embeds the entities from the *Repertory* into the National Library of Wales' wider knowledge graph. This means that any other data the library holds for people or places in their archives or visual collections becomes connected to the *Repertory* data.

60

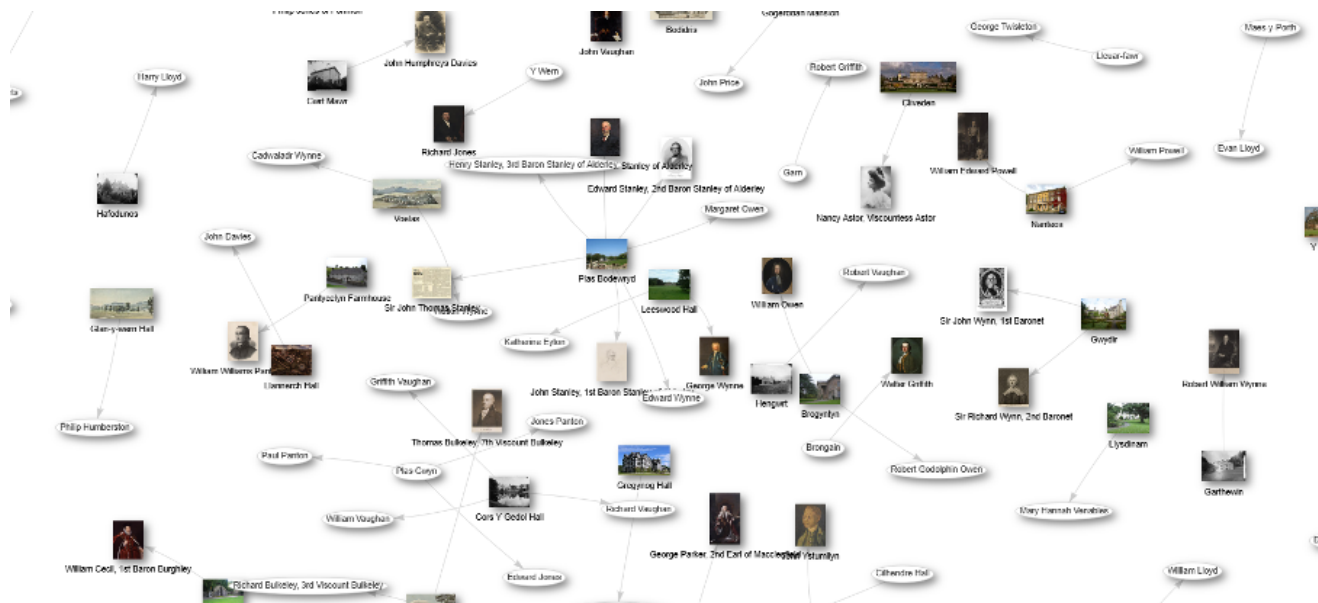


Figure 13. Graph showing people in NLW collections with some association to the places named in the *Repertory*.

3.6.4 Recording of potential further work on dataset

A list of potential enhancements or additions to the dataset which were out of scope for this stage of the project but could form part of future work packages were recorded as issues on the GitHub site. Examples include the addition of ISO language codes to manuscript records, the creation of a separate index for bibliographic references, and the manual encoding of relationships occurring within indexes (places referenced in person entries etc.).

61

3.7 Initial Research Findings

To conclude this section, it is worth underlining the research potential of the work. Besides the advantage of widening access, digitisation of the *Repertory* opens many new avenues, only some of which have been touched upon above. The simple ability to combine searches is powerful. At base, it makes it straightforward to find manuscripts in which the work of two specific authors is found, or to which two particular scribes contributed, or to combine categories and find, for example, manuscripts containing text X, copied by scribe Y and later in the collection of owner Z. These searches are possible using the print indexes but would be slow and laborious. The encoding of dates and the expansion of geographical information, however, add categories that cannot be accessed through the indexes and will allow, for example, a search on manuscripts dating before 1450, originating from Anglesey. Visualisation of such searches adds a further powerful element to an analytical tool that promises to transform the study of the development of Welsh written culture across space and time. Further extensions of the resource would then open new horizons. The ability to pursue searches across biographical material, across annotated editions of texts, or across borders into other collections – where comparative work might trace the influence of particular classical texts or authors, or of the manuscript contexts of particular types of text (such as legal, medical, religious or literary), will highlight countless connections and open many doorways to further research.

62

In practical terms, initial data analyses are proving helpful in one of the chief tasks of the *Repertory* – the identification of unknown scribes. Network analysis allows us to assemble the intellectual milieu of these scribes (indicated with identifiers beginning with X), which helps us both to position them in relation to other scribes in the dataset, and to highlight possible candidates for their identification. At present checking these hypotheses would entail going back to images of the manuscripts or to the manuscripts themselves, but the future integration of IIIF images of manuscript pages would allow this process to happen within the dataset. It is likely that in future a combination of data-driven analyses and AI-enabled handwriting comparison of IIIF images will be transformative in this important research area.

63

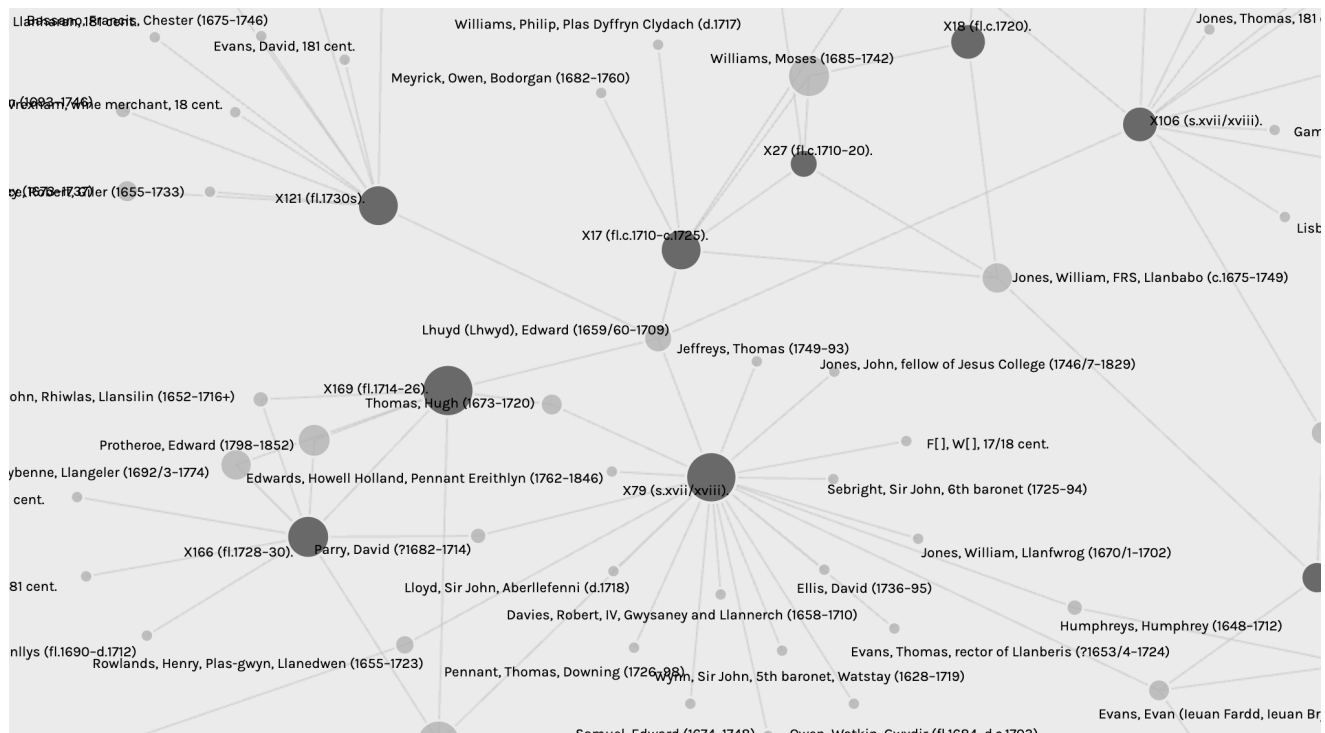


Figure 14. Unidentified scribes (identifiers beginning with X) related to an intellectual network centred on Edward Lhwyd.

4. METHODOLOGICAL CONCLUSIONS

Our project highlighted some important general issues relating to the conversion of print volumes to digital datasets:

Establish right at the outset if the conversion will be a one-off process or if it will need to be repeatable. This will have significant consequences for your workflow as it is very difficult to insert semi-automated or manual processes into repeatable conversions, because the manual adjustments will need to be made each time the conversion is run. This means that repeatable conversions often require extensive coding of edge cases, which is time consuming and can lead to bloated code. On the other hand, one-off conversions rely heavily on testing – and you will probably need to establish metrics and procedures to ensure that the process is doing what you expect at each stage. This, inevitably, will mean a lot of manual checking of outputs and exceptions. It is worth being very careful here – many conversions are envisaged as one-offs but end up needing to be re-done after manual editing of the source or the results. Communication between project participants is key – it is not uncommon for researchers to continue to edit the source data while the conversion is in process, which can lead to real problems in maintaining currency between versions. You should also ensure that plans for training or infrastructure requirements to allow researchers to edit the converted dataset are in place before the conversion takes place, to minimise the period where the dataset has to remain static.

As long as the source data is stable through the conversion process, a useful strategy is to retain the original text of each identified identity (a manuscript, a person, a work, etc.) in the output, in addition to the granular values created by data parsing. This will allow you to go back to the original source in the event of errors in the conversion process without having to rerun the whole process and will also enable iterative approaches to data parsing. In this context, it is not necessary to write one very long, unmaintainable script to do the whole conversion process at once. It can be better to use a series of scripts to extract specific data and test at each stage.

The assignment of unique identifiers to all entities is a vital part of the construction of most datasets. These should be prefixed with something which makes it clear what is being identified (e.g. manuscript_, person_, work_) but the identifier itself should be meaningless – in our case we simply used a running number for each type of entity. Meaningless identifiers have the advantage that they don't need to be changed if information about the entity changes (using shelfmarks as identifiers for manuscripts seems fine until they are reclassified, which happens more often than

you might think) and also that identifiers don't have to be reassigned in the event of the addition of more entities to the dataset (as would be the case if your identifiers reflected some kind of ordering of the entities).

In our experience, it is better to assign project-specific identifiers during the conversion process rather than attempting to match entities against external sources (Wikidata, VIAF, etc) on the fly. These local identifiers can then be reconciled against external identity providers as a separate stage. There are a number of advantages to this approach. Not all entities will be recorded in external identity schemes, and even for resources which allow you to create entries it is easier to do this as a separate process after the data conversion. Retaining your own identity records for entities in your dataset also allows you to record project-specific information about them which might not fit into a more general Wikidata or VIAF entry. Finally, the process of reconciliation against external identity providers can be complex and require further research (particularly in a dataset such as the *Repertory* where there is a high level of duplication of common names). Once you have established your own identity record for an entity you can either add a range of external identifiers to the record, or add the Wikidata identifier (or equivalent) and rely on that resource's aggregation of identifiers to network into the wider world of related datasets. The second approach is more efficient, but does create a dependency on an external data source.

68

Collaborative workflows were essential to the creation of the dataset, and will be equally important in its ongoing development and curation. On the technical side, GitHub was useful in a number of ways, from its support for collaborative editing to its versioning capabilities. We also made extensive use of GitHub issues to assign tasks during the conversion process and in building a to-do list for future stages of the project. It was felt to be essential that subject specialists with knowledge of Welsh manuscripts were able to directly engage with the data and with technical aspects of the project. We held training sessions on TEI, the Oxygen XML Editor and GitHub, and team members are now able to edit and manage the dataset independently in a secure environment. Too many projects have a sharp divide between technical and research activities, which can lead to massive problems with sustainability once the funded part of the project comes to an end, as those with responsibility for the long-term future of the data simply don't understand it.

69

Our choice of text encoding as a method and TEI as a data standard was fundamental to our approach – crucially retaining the idea of the *Repertory* as a text, rather than simply a source to be quarried for “useful” fragments of information. Flanders and Jannis point to a separation between data modelling and implementation of the model – that both logical and conceptual models should be “designed without any reference to a specific implementation, so that the implementation can be optimized or even entirely replaced with a different one at a later point in time” [Flanders and Jannidis 2016, 230]. This is certainly helpful in thinking through the separation between data and platforms – but implementations, here in the form of data standards, are epistemological frameworks with their own inherent methods and approaches. Kate Singer has emphasised the relationship between text encoding and engagement with the text – how the “... practice of encoding effectively engages the encoder in a determinative act of reading” [Singer 2013] – inviting the encoder to make assertions about the text within a defined set of parameters.

70

In essence, the conversion from print volumes to digital dataset was an exercise in modelling the implicit structures, connections and intentions of the *Repertory*. The most obvious modelling process was the pragmatic task of parsing the text into distinct values based on layout, format, order and textual content. But intertwined with this was a deeper layer of conceptual modelling – what are the purposes of the *Repertory*? How are these reflected in its structures and mappings? More than the technical implementation of the data conversion, it was this close reading of the *Repertory* which lay at the heart of the project.

71

This “decoding/encoding” of the *Repertory* was a collaborative process involving a team with different backgrounds and areas of expertise, from Python programming to linked data to manuscript studies to those in between – the “hybrid people” with a foot in both camps who are so important in the intersection of the humanities and the digital [Siemens et al. 2011]. The use of standard approaches such as TEI and Wikidata links the *Repertory* to wider networks concerned with the representation and use of cultural heritage materials in a digital environment. The ongoing conversation amongst the diverse community of people with an interest in the *Repertory* specifically and manuscript datasets more generally is arguably the most vital factor in its future development and sustainability.

72

One final methodological conclusion is that in technical work of this type it is useful to have some kind of metric or framework against which to measure your progress, even if you find good reasons to adapt or ignore some aspects of it. For us, the clear and simple guidelines in sections 1 and 2 of the principles established by the Endings Project [Endings Project 2021] were a good reference point in the construction and documentation of the dataset. We managed to comply with all of the principles except for point 2.2 (licensing) which will form part of further discussions around the open release of the data later in 2026. Sections 3–5 were also instrumental in our thinking around the minimal computing approaches to platforms discussed in the next section.

73

5. NEXT STEPS

Now we have the dataset, what do we do with it? In this section we will outline possible next steps for the *Repertory*, and in doing so touch on some broader conceptual issues relating to datasets emerging from collections. Taken as a whole, these thoughts can be usefully classified as an approach which takes practical measures to facilitate greater access to and engagement with infrastructure and resources, with the aim of increasing the diversity of people who are able to engage directly in both the creation and the use of datasets – particularly in the context of minority cultures and languages. Here we draw on the “full stack” framework proposed by Alan Liu [Liu 2020] and projects such as Full Stack Feminism [Bassett, Jarrett, and Webb 2025]. Talking about the “big tent” model of diversity in a Digital Humanities context, Liu points to a “glaring mismatch” between ideas of diversity and the technical infrastructures which underpin work in this area:

74

Big tent named the DH diversity platform as ideology. But it contributed no guidance at all for it as a technical platform. Putting this more carefully: what is at stake is not a binary difference between ideology and technology but instead a not yet-understood need in the field for a fused techno-ideological apparatus – a platform in all senses – that can do urgently needed work

[Liu 2020, 133].

We will now look in more detail at four core areas of “urgently needed work” which will be central to the next stages of our project. One of the most striking aspects of this thinking is how often there are parallels between our work on the dataset and the manuscript culture it describes.

75

5.1 Data – sustainability and accessibility

The sustainability of data and interfaces is one of the chief problems in collections-based digital work – especially in a system dominated by short-term funding cycles and institutional cutbacks (see e.g. [Smithies et al. 2019], [Cummings 2023]). In this context we have adopted an approach where there is a clear separation between the most valuable long-term asset of the project (the data) and web-based interfaces to the data (platforms). The data is stored in TEI, a widely-used and well-documented open standard. As a text-based format, TEI does not require a system or database for either storage or use, and is extremely lightweight – the whole *Repertory* dataset currently stands at 26.3 MB. This, combined with the huge amount of detailed documentation available on TEI (and the many hundreds of projects which use it) mean that it is very likely that in ten- or twenty-years time scholars would be able to access and make use of the data without the need for specific tools or knowledge. TEI’s use of the almost ubiquitous XML data format in addition to the large amount of open-source development which has been done around it mean that it is probable that those scholars would also be able to start surfacing, investigating and analysing the data without the need for expensive software development. The long-term future of the data at least can be secured at minimal or no cost.

76

There are bigger questions around how and where the data is stored and made available. Currently the dataset is on a private GitHub site where we take advantage of GitHub’s versioning capabilities and support for collaborative editing workflows. GitHub is currently free at the point of use, but it is owned by Microsoft, and history teaches us that large and well-established services under private ownership are subject to fundamental changes in their operation or indeed complete shutdown – Skype being just one recent example. One option for long-term storage of the data would be to use institutional repositories such as Cambridge University’s Apollo – a good solution for this project, but one which relies on the kind of affiliation which could be a barrier to participation for others. Another option would be to deposit the

77

data in a large non-profit data repository such as Zenodo – which also would bring benefits in terms of referencing the dataset (see 5.3 below). It may be that long-term sustainability and access is best served by a mixed approach, with multiple copies of the data distributed across different repositories. This would have the pleasing effect of mirroring the manuscript tradition which the data describes, though it would also be susceptible to similar problems of divergence and variation in the longer term.

One beneficial outcome of our focus on creating and storing the data in a standard, open format is the potential for audiences and collaborators to work with it directly, rather than simply accessing it through platforms and interfaces. A major theme in the current stage of the project was the breaking down of barriers between project members from differing backgrounds and with different skills and knowledge – opening the way to the kind of “fused techno-ideological apparatus” which Liu calls for. We envisage a model where direct interaction with data does not require institutional affiliation, purchase of software or access to expensive infrastructure – but does lead to the dissemination of skills and knowledge transfer between “technical” and “academic” aspects of projects. This “democratisation” of the data brings the idea of the open stack out of the realm of physical and virtual infrastructure, and into the cultures which exist in the interfaces between research and technology.

78

5.2 Linking and integration

As detailed in section 2, the *Repertory* is situated in a network of resources, both online and print, including digitised images, digital editions, biographical and geographical datasets, authority and identity schemes, and, indeed, other manuscript catalogues. The *Repertory* is central to this network, but does not stand alone. As Jacqueline Burek points out:

79

... potential researchers should be aware that, if they wish to investigate Welsh culture beyond the bounds of the ‘Welsh tradition’, they should use the *Repertory* alongside other resources (many of which Huws himself suggests)

[Burek 2023, 462].

There are two strategies for connecting the *Repertory* to its intellectual hinterland, each with its advantages and disadvantages. The first might be described as the “linking” approach, in which resources are kept as separate entities and are connected through the use of common identifiers. As described in section 4.4, the *Repertory* makes extensive use of identifiers for its internal connections, but these are currently local to the dataset. Datasets such as FIHRIST commonly use external identifiers derived from authority schemes such as VIAF, Getty and the Library of Congress to connect with other resources, but contribution to these schemes is dependent on institutional affiliation. Even identity providers with a more open stance on contribution, such as Wikidata, have policies about the “notability” of entries which can mitigate against minority languages and cultures.

80

To get around these problems, The National Library of Wales has made innovative use of a local copy of Wikibase, the software behind Wikidata, which is connected to but not dependent on the main Wikidata dataset. The extension of a similar solution across manuscript datasets would be a fruitful avenue to explore. In addition to providing common identifiers for elements such as names and places, this approach would also open the possibility of creating controlled vocabularies for aspects of manuscript description such as script, material, binding etc. which are conceptual rather than language-focussed and allow us to explore manuscript cultures across philological boundaries. This work would build on our existing activity in developing a consolidated TEI schema for special collections material.

81

The linking approach has many advantages – data compatibility, reducing redundancy, enabling the cross-analysis of datasets and retaining the sense of each data source as a separate entity with its own provenance and rationale. However, it can raise problems in terms of sustainability, creating dependency on the continued existence of external sources (what James Dobson refers to as “text-external referential systems” [Dobson 2019, 74]) in order to make sense of the data. Our scholars who wish to use the data in twenty years time might find it full of identifiers which refer to other resources and datasets which are no longer extant.

82

An alternative approach, the “integration” method, would involve actually *including* information from external sources in the *Repertory* itself – whether it be transcriptions derived from digital editions, biographical and geographical information, or extra description from other catalogues. This approach creates self-contained datasets which are in some senses more sustainable, but bring different problems in terms of maintaining the currency of the same data across several resources. It also relies heavily on the recording of data provenance – the origin of each part of the data – which we will discuss in more detail in the next section.

83

5.3 Additions, edits and referencing

In section 2.3 we talk about the tension between the *Repertory* as a work with its own purposes and internal logic, and the idea that it openly invites additions and edits. Changes to the dataset, whether they be derived from existing resources (see above) or from new research, require a detailed approach to data provenance. In our previous work on Islamicate manuscripts, we emphasised the importance of “... adopting a transparent and non-hierarchical model of recording contributions directly into the data, whether they be by people or by processes, and taking advantage of the versioning capabilities of modern repositories to record the various stages in the development of our datasets” [Faghihi and Jones 2024], arguing that without such a model it becomes impossible to interpret the outputs of our analyses and investigations. Again, we are reminded of the materials of the *Repertory* – the additions, amendments, comments and reworkings (some named, some not) in the manuscript culture which is its object of study. The prime purpose of the *Repertory* is to work out who wrote what and when – we don’t want to leave the same challenge to future users of the dataset.

84

TEI provides its own mechanisms for the recording of data provenance, allowing for the recording of the overall sources of a manuscript description through the <listBibl> and <recordHist> elements, combined with a granular approach to noting the provenance of particular parts of record using through the use of <revisionDesc> and the change attribute and element. FIHRIST takes an interesting approach by including editors of the manuscript descriptions in its name index, allowing for the retrieval, for instance, of all of the manuscripts added or edited by a particular person. This is not just a convenience for reviewing edits to the data – it is also a deliberate step towards acknowledging the labour which goes into manuscript research. Things become a little more complicated when automated processes are applied to the data. These can be recorded using <revisionDesc> and <change> in the same way as manual edits, and we have adopted a policy of explicitly linking to the code through which the changes were made on open repositories. The use of AI for editing purposes is likely to introduce further complications in this area in the future.

85

TEI also provides a method for establishing the certainty (or lack of certainty) of an assertion. For manual edits, this might be based on the availability of evidence – for automated procedures it could reflect either statistical likelihoods from quantitative methods, or (for AI or similar) a general sense of how much we “trust the machine”. Flanders and Jannidis point to the recording of uncertainty as a general problem in humanities data modelling, and as a specific issue for TEI, where it is “... implemented unevenly: uncertainty about details of the transcription is modeled in one way, while uncertainty about the chosen markup is modeled in another, and there are some aspects of uncertainty that cannot yet be modeled effectively ...” [Flanders and Jannidis 2016, 235]. The application of multiple layers of automation to datasets, with evidence for new processes being directly drawn from the outputs of previous ones, makes strategies for recording the provenance and certainty of data increasingly important, and in future stages of the project we will need to do work in this area.

86

While TEI gives us a range of granular methods for recording data revisions and provenance *within* a record, it does not solve the problem of how to provide a reliable mechanism for scholars to reference particular versions of the dataset as a whole. A useful model here is the integration between GitHub and Zenodo, which allows for the creation of labelled releases of data which are then assigned a DOI for reliable reference to a particular version of the data. This does rely on a consistent approach to release schedules, and also on an editorial decision on what constitutes a “release” of the data (regular releases over time, or releases triggered by major changes to the dataset).

87

5.4 Platform

Perhaps the most problematic aspects of collection-driven digital humanities projects are connected to the platforms which surface them. These have tended to rely on expensive web development, institutional infrastructure and specific technical skills for both their creation and their maintenance. There has arguably been an over complication in terms of user experience, with key functions such as search and display intertwined with features such as visualisation and analytical tools which are more difficult to maintain. Security concerns lead to the need for regular updates to software and adjustments to code which, again, rely on technical knowledge and infrastructure. The net result is that platforms tend to be run by institutions with access to infrastructure and staff with technical skills, and even in these cases they often don't outlast the period in which the project is being directly funded. The cultural result is that platforms and datasets tend to follow the interests, assumptions and worldview of these institutions.

88

The most radical approach to this problem might be termed the “no platform” solution – where the output of the project is openly-available, standard, well-documented data and the process of accessing and analysing that data is seen as an ongoing process which will inevitably involve multiple agents, producing necessarily ephemeral interfaces which are directed towards specific purposes and views on the dataset. In a sense this is where we are with the *Repertory* – the dataset is there, and we have produced various views on it including maps, graphs and network analyses which provide new views on the data but are not expected to last. But we would like a platform!

89

The second approach might be termed the “minimal computing” solution. Here, we identify the core features which users are likely to require from an interface to the data (display, search, faceting, download) and provide the simplest possible platform which will fulfill these functions – while also making the data available for more complex uses such as visualisation and analysis. In this model we would (again) follow the guidelines provided by the Endings Project [Endings Project 2021] around the use of static HTML for display and standard Javascript for search and faceting. The staticSearch project has already made considerable progress in this area. A solution based around static HTML would dramatically reduce reliance on infrastructure (being very lightweight to host) and technical expertise (static HTML can be generated through comparatively simple Python or XSLT scripts) as well as alleviating concerns around security (there is no code on the server side to be hacked).

90

5.5 Summary

What would the effect be of removing or reducing institutional or financial barriers to access to the point at which a much broader and more diverse range of people would be able to directly contribute to work of this kind? We hope that in addition to addressing real concerns around sustainability and access, work of this kind would reduce or remove the need for institutional patronage, and therefore foreground a different set of priorities and interests – or at least open up that space to more experimental ideas and approaches. This could include a less siloed approach to the data, multilingual and cross-cultural approaches, and the kinds of inductive research which are promised by large open datasets – and which might illustrate the connectedness of cultures, concepts and epistemologies which are usually considered separately. To leave the last word to Alan Liu:

91

So here is my program for the digital humanities. I hope digital humanists can build more experimental, cross-domain corpora designed on purpose to be other than tidy. Digital humanists should make corpora that mix disciplines, provenances, formats, metadata structures, and so on to remix the evidence of the human

[Liu 2016].

The first stage of “remixing” the *Repertory* is complete – and we look forward to embarking on the kind of interdisciplinary, experimental, collaborative work which Liu argues for, and which could genuinely tell us new things about manuscript cultures in Wales and beyond.

92

Works Cited

Baker, Salway, and Roman 2022 Baker, J., Salway, A., and Roman, C. (2022) “Detecting and characterising transmission from legacy collection catalogues”, *Digital Humanities Quarterly*, 16(2). Available at: <https://www.digitalhumanities.org/vol/16/2/000615/000615.html>.

- Bassett, Jarrett, and Webb 2025** Bassett, C, Jarrett, K. and Webb, S. (2025) "Stacks and intersections: Feminist thinking in digital humanities, a view from 'these islands'", *Digital Humanities Quarterly*, 19(3). Available at: <https://dhq.digitalhumanities.org/vol/19/3/000807/000807.html>.
- Burek 2023** Burek, J.M. (2023) "A repertory of Welsh manuscripts and scribes, c. 800–c. 1800 by Daniel Huws (review)", *Manuscript Studies*, 8(2), pp. 458–462. Available at: <https://muse.jhu.edu/pub/56/article/916144>.
- Cummings 2023** Cummings, J. (2023) "Academics retire and servers die: Adventures in the hosting and storage of digital humanities projects", *Digital Scholarship in the Humanities*, 17(1). Available at: <https://digitalhumanities.org/vol/17/1/000669/000669.html>.
- Dobson 2019** Dobson, James E. (2019) *Critical Digital Humanities: the search for a methodology*. Champaign: University of Illinois Press.
- Endings Project 2021** Endings Project (2021) *Endings principles for digital longevity, version 2.1*. Available at: <https://endings.uvic.ca/principles.html>.
- Faghihi and Jones 2024** Faghihi, Y. and Jones, H. (2024) "Manuscript catalogues as data for research: from provenance to data decolonisation", *Digital Humanities Quarterly*, 18(3). Available at: <https://www.digitalhumanities.org/vol/18/3/000751/000751.html>.
- Flanders and Jannidis 2016** Flanders, J. and Jannidis, F. (2016) "Data modelling", in Schriebman, S., Siemens, R. and Unsworth, J. (eds) *A new companion to digital humanities*. Chichester: Wiley Blackwell, pp. 229–237.
- Formigatti 2017** Formigatti, C.A. (2017) "Sanskrit manuscripts in the Cambridge University Library: Three centuries of history and preservation", in Vergiani, V., Cuneo, D., and Formigatti, C. (eds.) *Indic manuscript cultures through the ages: Material, textual, and historical investigations*. Boston, MA: De Gruyter, pp. 3–46.
- Guy 2024** Guy, B. (2024) "A new era for Welsh manuscript studies", *The Library*, 25(1), pp. 88–95. Available at: <https://doi.org/10.1093/library/fpae001>.
- Huws 2022** Huws, D. (2022) *A repertory of welsh manuscript and scribes, c.800–c.1800*. Aberystwyth: The National Library of Wales and the University of Wales Centre for Advanced Welsh and Celtic Studies.
- Liu 2016** Liu, A. (2016) "N + 1: A Plea for Cross-Domain Data in the Digital Humanities", in *Debates in the Digital Humanities 2016*. Minneapolis, MN: University of Minnesota Press. Available at: <https://dhdebates.gc.cuny.edu/read/untitled/section/d7f3fec8-4b39-4269-91c5-536a9bf25355>.
- Liu 2020** Liu, A. (2020) "Toward a diversity stack: Digital humanities and diversity as technical problem", *PMLA/Publications of the Modern Language Association of America*, 135(1):130–151. Available at: <https://doi.org/10.1632/pmla.2020.135.1.130>.
- Russell 2023** Russell, P. (2023) "In the penumbra of the Repertory: medieval manuscripts from Wales", *Studia Celtica*, 57(1), pp. 151–164(14). Available at: <https://doi.org/10.16922/SC.57.8>.
- Siemens et al. 2011** Siemens, L., Cunningham, R., Duff, W., Warwick, C. (2011) "A tale of two cities: implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities", *Literary and Linguistic Computing*, 26(3), pp. 335–348. Available at: <https://doi.org/10.1093/lilc/fqr028>.
- Singer 2013** Singer, K. (2013) "Digital close reading: TEI for teaching poetic vocabularies", *The Journal of Interactive Technology and Pedagogy*, 3. Available at: <https://jitp.commons.gc.cuny.edu/digital-close-reading-tei-for-teaching-poetic-vocabularies/>.
- Sloan and Nyhan 2021** Sloan, K. and Nyhan, J. (2021) "Enlightenment architectures: the reconstruction of Sir Hans Sloane's cabinets of 'Miscellanies'", *Journal of the History of Collections*, 33(2), pp. 199–218. Available at: <https://doi.org/10.1093/jhc/fhaa034>.
- Smithies et al. 2019** Smithies, J., Westling, C., Sichani, A., Mellen, P. and Ciula, A. (2019) "Managing 100 digital humanities projects: Digital scholarship & archiving in King's digital lab", *Digital Humanities Quarterly*, 13(1). Available at: <https://digitalhumanities.org/vol/13/1/000411/000411.html>.

Recommendations

DHQ is testing out three new article recommendation methods! Please explore the links below to find articles that are related in different ways to the one you just read. We are interested in how these methods work for readers—if you would like to share feedback with us, please complete our short evaluation survey. You can also visit our documentation for these recommendation methods to learn more.

SPECTER Recommendations

Below are article recommendations generated by the SPECTER model:

1. Encoding for Endangered Tibetan Texts, 2007, Linda E. Patrik, Department of Philosophy, Union College
2. e-Science for Medievalists: Options, Challenges, Solutions and Opportunities, 2010, Peter Ainsworth, Dept of French and Humanities Research Institute, University of Sheffield; Michael Meredith, Humanities Research Institute, University of Sheffield
3. The Lives of Mistresses and Maids: Editing Victorian Correspondence with Genealogy, Prosopography, and the TEI, 2022, Kailey Fukushima, University of British Columbia; Karen Bourrier, University of Calgary; Janice Parker, Calgary Public Library
4. The Ends of Editing, 2009, Peter M. W. Robinson, University of Birmingham
5. "May the Text Rise up to Meet You": New Ways of Reading Old Manuscripts, 2009, Eugene Lyman, University of Rhode Island

DHQ Keyword Recommendations

Below are article recommendations generated by DHQ Keywords:

1. Manuscript Catalogues as Data for Research: From Provenance to Data Decolonisation, 2024, Huw Jones, Cambridge University Library; Cambridge Digital Humanities; Yasmin Faghihi, Cambridge University Library; Cambridge Digital Humanities
2. Reading Chicago Reading: Quantitative Analysis of a Repeating Literary Program, 2020, John Shanahan, DePaul University; Robin Burke, University of Colorado — Boulder; Ana Lučić, Beckman Institute, University of Illinois
3. Mining Embodied Emotions: A Comparative Analysis of Sentiment and Emotion in Dutch Texts, 1600-1800., 2018, Inger Leemans, Faculty of Humanities, Vrije Universiteit Amsterdam, The Netherlands; Janneke M. van der Zwaan, Netherlands eScience Center, Amsterdam, The Netherlands; Isa Maks, Faculty of Humanities, Vrije Universiteit Amsterdam, The Netherlands; Erika Kuijpers, Faculty of Humanities, Vrije Universiteit, Amsterdam, The Netherlands; Kristine Steenbergh, Faculty of Humanities, Vrije Universiteit, Amsterdam, The Netherlands
4. The Design of an International Social Media Event: A Day in the Life of the Digital Humanities, 2012, Geoffrey Rockwell, University of Alberta; Peter Organisciak, University of Illinois, Urbana-Champaign; Megan Meredith-Lobay, University of Alberta; Kamal Ranaweera, University of Alberta; Stan Ruecker, Illinois Institute of Technology; Julianne Nyhan, University College London
5. Towards 3D Scholarly Editions: The Battle of Mount Street Bridge, 2019, Costas Papadopoulos, Maastricht University; Susan Schreibman, Maastricht University

TF-IDF Recommendations

Below are article recommendations generated by the TF-IDF Model:

1. Recovering the London Stage Information Bank: Lessons from an Early Humanities Computing Project, 2017, Mattie Burkert, Utah State University
2. Conjectural Criticism: Computing Past and Future Texts, 2010, Kari Kraus, College of Information Studies and Department of English, University of Maryland
3. Manuscript Catalogues as Data for Research: From Provenance to Data Decolonisation, 2024, Huw Jones, Cambridge University Library; Cambridge Digital Humanities; Yasmin Faghihi, Cambridge University Library; Cambridge Digital Humanities
4. The MoEML Pedagogical Partnership Program, 2017, Janelle Jenstad, University of Victoria; Kim McLean-Fiander, University of Victoria; Kathryn R. McPherson, Utah Valley University
5. Decolonial by Design: Building Sekuru's Stories, 2024, Jennifer W. Kyker, University of Rochester



This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.