Building The Human Remains Digital Library (HRDL)

The editorial steps behind sourcing, cleaning, and aligning a cross-discipline multi-lingual thematic resource

RUTH NUGENT
University of Liverpool, ruth.nugent@liverpool.ac.uk

JAMES O. BUTLER
University of Liverpool, james.butler@liverpool.ac.uk

GLENN CAHILLY-BRETZIN
University of Liverpool

ADAM DAUBNEY
University of Liverpool

THOMAS J. FARROW*
University of Liverpool

KATHERINE FOSTER
University of Liverpool

ISABELLE GRIBOMONT
University of Louvain

LLEWELYN HOPWOOD
Cardiff University

Just Accepted

Abstract

The Human Remains Digital Library (HRDL), funded by a seven-year UKRI Future Leader's Fellowship, is the first cross-disciplinary, multilingual thematic resource created to support researchers, heritage professionals, and decision makers managing historic human remains encountered in British churches and cathedrals. It integrates for the first time curated excerpts from seventh- to nineteenth-century written accounts (whether real, embellished, or imagined) with standardized summaries of modern archaeological evidence for historic exhumations, innovating both temporal–spatial analysis of burial (re-)locations and investigation of evolving ethical, emotional, and theological notions of 'respect' for the dead. HRDL departs from manuscript-centric libraries by integrating multilingual textual excerpts and standardized archaeological summaries within a single thematic, cross-period corpus, enabling site- and genre-specific retrieval and analysis, advancing digital library practice through a combined workflow: discovery via concordance- and regex-driven multilingual search; bespoke re-OCR and cleaning pipelines; cross-period spelling normalization; controlled vocabularies bridging historical text and archaeological terminologies; and a CIDOC-CRM–aligned metadata ontology served through a user-centric Blazor interface. The resulting corpus (3,294 records at launch of over a million words, across multiple sub-corpora and languages, with new translations for previously inaccessible medieval sources) provides site-specific and period-specific evidence to support church building maintenance and (re-)burial policy while opening large-scale research questions previously impossible to address. This article outlines the editorial

* also known as Thomas J. Fitzgerald

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1556-4711/2026/5-ART
https://doi.org/10.1145/3813803

and technical rationale behind HRDL's design and contributes a reproducible blueprint for building thematic, multilingual, cross-evidential digital libraries aimed at both academic analysis and applied knowledge exchange.

CCS CONCEPTS •Information systems → Data management systems; Digital libraries and archives; • Applied computing → Document management and text processing; Document capture; • Applied computing → Enterprise computing → Enterprise ontologies, taxonomies and vocabularies;

Additional Keywords and Phrases: digital libraries, church archaeology, digital content challenges, standardisation, linked data, thematic builds

1 INTRODUCTION

Commencing in November 2019, the aim of *The Human Remains: a digital library of historic British mortuary science and investigation project* is to support historic cathedrals and churches in Britain who manage human remains inside their buildings. Burials inside British churches and cathedrals were largely banned by the Burial Acts of the 1850s, so new interments are rare in the 20th and 21st centuries. People continued to be buried in churchyards or cemeteries instead. However, cathedrals and churches pre-dating the Burial Acts often reveal historic human remains during maintenance and restoration of the building; in building extensions; adding disability access or fire safety features, and the like. Gaps in — or the absence of — appropriate policies at national level (England, Scotland, Wales) and between church denominations; complex logistical issues for removing and reburying or storing human remains; and ethical concerns surrounding 'respect' for the dead are an ongoing issue today, in churches, as well as in museums and archaeology generally [1, 2]. This enduring challenge motivates the project's dual focus on actionable burial management and historically grounded, ethics-aware engagement with human remains. However, churches in Britain have particularly pronounced issues given their ongoing use as burial grounds, as well as managing historic burials during maintenance and rebuilding projects while maintaining their role as active places of worship and community. Unlike museums, churches must often carefully navigate theology, requirements of the clergy, church heritage standards, and local congregations' expectations when human remains are exposed. This intersection of living worship, heritage stewardship, and mortuary evidence underscores a distinctive stakeholder landscape that existing datasets and tools have not adequately served.

Yet discussions addressing these issues in museums have tended to overlook churches, both individually and collectively, which generally lack research capacity and support from academic communities. A central barrier is the absence of quantifiable, accessible, searchable corpus tracing where human remains have been (re)located over the generations inside the church buildings, and how people in the past expected their remains to be treated respectfully, which varied over time and context. HRDL is designed explicitly to fill this gap by making dispersed, multi-language evidence open access, discoverable, and comparable, enabling the first large-scale analysis of historic trends and priorities which can inform modern policies.

Church sites have complex, often idiosyncratic, histories of burials being disturbed by national and local events, whether through fire, flood, riots, rebuilding, collapse, re-ordering, maintenance, religious and political upheavals, local feuds, or other issues. Limited archaeological excavation in active churches often constrains the capacity to record, maintain, catalogue, and otherwise make available archaeological findings relevant information, which is beyond the scope of this article to detail [3]. No single solution or complete timeline of all burial disturbances or the decisions behind it is feasible. A 'Big Data' synthesis of existing historical and archaeological records is therefore essential to reveal robust temporal-spatial patterns in burial management and to trace shifting social-theological constructions of 'respect'.

As a result, *The Human Remains* project created a digital library (HRDL) for use by researchers, practitioners, and stakeholders needing to identify: (a) patterns of burial management over time; and (b) period-specific debates concerning the ethical, emotional, and theological issues of exhuming and relocating human remains from Christian contexts. Working beyond just the language of such accounts, we can analyse the material conceptually to detect historic patterns, trends, and shifts. The project could not,

Just Accepted

therefore, simply replicate relational structures or interfaces from other digital libraries for historical texts since HRDL was not aimed at annotating manuscripts, transcribing handwritten documents, or creating a complex PDF-viewer. Instead, HRDL provides users with site-specific, period-specific, genre-specific evidence from evidence drawn from excerpted historical accounts and distilled archaeological reports spanning the 7th-19th centuries. Complementing this textual material with archaeological evidence meant distilling lengthy excavation reports into key information of historic exhumations. Thus, HRDL departs from manuscript-centric libraries by integrating multilingual textual excerpts and standardized archaeological summaries within a single thematic, cross-period corpus, enabling site- and genre-specific retrieval and analysis. Achieving this required an end-to-end editorial pipeline—multilingual discovery using concordance/regex; re-OCR and cleaning; spelling normalization; controlled vocabularies; and a CIDOC-CRM–aligned metadata model — implemented within a user-centric Blazor interface.

This meant overcoming substantial editorial and technical challenges to create HRDL, which are outlined in this article. The aim is two-fold: (1) to provide a clear rationale for producing a digital library of this nature within the current landscape of digital libraries for historical sources; and (2) offer a transparent model of editorial decisions, design principles, and technical solutions to support other Big Data projects looking to compile equally complex libraries in a creative manner, and/or work with a variety of end-users, who do not have an obvious model to draw upon. We outline the challenges faced, and solutions developed, by *The Human Remains* project to design a digital library aimed at a range of academic and non-academic audiences who have different priorities for the same material, and contribute four areas of addressing these: (1) a replicable workflow for thematic, multilingual corpus compilation across twelve centuries; (2) integration of historical texts and archaeological grey literature under shared, controlled vocabularies; (3) a bespoke, interoperable ontology mapping Dublin Core to CIDOC-CRM for content-level metadata; and (4) an accessible User Interface (UI) that combines full-text/metadata search with discovery-led filtering for cross-sector users.

We begin with an overview of HRDL's requirements, challenges, and user needs. Each area of HRDL's development to meet these aspects is then explained chronologically, from the initial workflow to the library's current status. The workflow is discussed in three key stages: (1) creating the corpus; (2) structuring the library; (3) design principles for the interface and search facility. This required the adoption of approaches rarely employed in other digital libraries, which this article will expand on: (a) developing a thematic corpus without antecedent; (b) the challenges and editorial decisions for designing a multilingual, cross-disciplinary, cross-period corpus; (c) developing controlled vocabularies to standardise terminology; (d) the rationale and ontological framework for the library's underlying relational structure; (e) combining full-text and metadata searching with static category filtering. Collectively, these design choices demonstrate how thematic digital libraries can serve knowledge exchange beyond academia while maintaining the analytical rigour required for large-scale research.

Having demonstrated the opportunities and technical challenges for HRDL, these are then contextualised and contrasted with other digital libraries operating in a similar research space, examining the distinct challenges for a multilingual, cross-period, multi-evidential library aimed at facilitating both knowledge exchange and research. We then discuss how the design principles and lessons learned through this project might help others developing similar multi-period, multilingual, thematically-structured digital libraries aimed at non-academic audiences.

2 CREATING THE CORPUS

HRDL currently houses 3,337 records pertaining to historical and archaeological accounts of exhumation and burial management in churches and Christian contexts in Britain (England, Scotland, and Wales), predominantly from the 7th to 19th centuries AD (with some important early 20th century inclusions). This period spans the earliest written accounts of exhumation in Britain until Britain banned intramural burial (i.e. interment inside church buildings) to all but a rare few by the late 19th century. Most of the relevant sources for HRDL are not under copyright due to this date range, but provisions for quick takedown notices were made in case of any unanticipated claims. It also allowed the project to make some

Just Accepted

texts (or extracts of texts) accessible in a digital, typed, searchable format for the first time for research. Furthermore, HRDL also contains modern English translations of hitherto untranslated medieval texts (or extracts) in Latin, Old English, Middle English, as well as multi-period Welsh-language sources. HRDL is also a rare combination of historical texts and modern archaeological data. This far exceeded the initial number of records that was anticipated at the outset, where around a thousand were expected.

Table 1: Number of records per sub-corpora (at time of publication, live and embargoed material)

Sub-corpora	Date Range	Total Records
Archaeology	1900-2024	264
Cathedrals	1631-1929	136
Charnel	1500-1900	1120
Postmedieval	1631-1921	185
Royals	600s-1916	331
Saints	700s-1499	1264
Welsh	1060-1899	37
		3337

No constraints were placed on the genre or authorship of the texts for the digital library, as representativeness was not a factor in HRDL's design. The project prioritised already digitised sources, supplemented by occasionally digitising physical texts. No compilations or lists of examples existed of church burials from which the team could quickly generate records for the library. No single author, genre, archive, language, or period defines the library; rather, HRDL is defined by a *theme* of exhumation and burial management in historic British cathedrals and churches. Thematic digital libraries are rare, although see [4] for a thematic approach to multilingual eastern European datasets. As exhumation was often more common and more commonly reported at cathedrals, due to their higher-status, longevity, bureaucracy, influence on parish churches, and as sought-after sites for saints' remains [3], cathedrals were *foci* for HRDL.

Based on Nugent's doctoral work [5], the proposed corpus was divided into sub-corpora (per **Table 1**, above) to capture: 'Saints' and 'Royals' (individuals frequently exhumed and reported, especially in the medieval period); 'Charnel' (exhumed and stored mass collections of human bones); 'Cathedrals' (focussing on the surge of cathedral records and writings after the 16th century); 'Postmedieval' (covering all other printed materials from 16th century onwards); 'Welsh', 'Middle English' and 'Old English' (requiring specialist language skills, dictionaries, and archives); and 'Archaeology' (evidence of historic exhumation in 20th and 21st century cathedral excavation reports).

HRDL end-users also contributed unique texts for HRDL which directly informed both applied practices and outreach events. For example, the 19th century diary of James Frater was digitised for Chester Cathedral, which detailed Frater's discovery of human remains during his restoration work. Making it machine-readable meant the cathedral's maintenance team could search Frater's entries to avoid disturbing human remains during routine work. The use of HRDL's material to highlight areas of buildings likely to contain unrecorded instances of human remains is a wider critical research priority of the project. Similarly, a talk by the Custodian of St Magnus Cathedral Orkney in the late 20th century, Albert Thomson, which outlined his knowledge of the cathedral's many reburials including its two medieval saints, was digitised and is the focus of a forthcoming exhibition.

2.1 Medieval/Postmedieval divide

Medieval exhumations were represented mainly by saints and royals as these were most frequently reported at the time. An exhaustive search of all medieval sources for other types of exhumation (e.g. clergy, military, or laity) was not possible, even with a large team, budget, and project lifespan, although Knights Hospitaller and Templar reburials in Britain were ruled out by project collaborators due to lack of evidence [6].

With shifts in church records, burial strategy, and religious beliefs during the Reformation in the 16th century, as well as the ever-increasing mass of printed texts being produced, the postmedieval period

Just Accepted

(16th–19th centuries) was separated out from the medieval period (7th–15th centuries, per our specified range). The projects' medieval/postmedieval divide also reflects period divisions inherent in the digital repositories and online dictionaries consulted [7] which mapped onto to the languages and periods of the sub-corpora.

In reality the medieval/postmedieval divide is an arbitrary, hotly-debated, shifting boundary, with as much continuity as change during the 15th and 16th centuries where the boundary is said to lie [8–10]. The project was highly conscious of this issue and does not recognise a hard division between 'medieval' and 'postmedieval' on a specific date. However, feedback from HRDL beta users indicated that being able to separate 'medieval' from 'postmedieval' texts would be helpful for period-specific research interests or public engagement events, as long as the date(-range) of the records was clear. This is particularly useful for churches and cathedrals which changed status in the 16th century Reformation. It was a well-discussed temporal division in Britain historically [9] and is still used today in the categorisation of archaeological evidence [11]. The date ranges from the Portable Antiquities Scheme (PAS) were therefore applied to divide 'medieval' (AD410-1539) from 'postmedieval' (AD1500–1900) and from 'modern' (AD1901 onwards) so both textual and archaeological evidence could be grouped temporally in the same way for HRDL [11]. As the Reformation process began in the early decades of the 16th century, (hence the 1500–1539 overlap in the PAS schema), texts written or published in 1500–1539 were collated under 'Postmedieval' unless they clearly related to the topics in the medieval sub-corpora, which meant the thematic groupings often transcended a strict medieval/postmedieval divide.

Royals continued to be exhumed after the Reformation and a number of saints continued to be discovered into the 19th century. However, as the postmedieval period saw a surge in printing and literacy, reflected in numerous dedicated online archives for this period, so the volume of texts concerning exhumation grew in number, scope, variety, and often in length. Charnelling; church tomb openings; essays on church burial management and sanitation; guidebook anecdotes; antiquarian reports; restoration projects, and newspaper reports of 'ancient' human remains in churches tended to dominate the postmedieval collection, making it more multifaceted than the medieval corpora which generally focussed on the 'special dead'.

However, postmedieval texts could be reprinted medieval material or duplicated verbatim reports. Cross-referencing between sub-corpora was therefore imperative. Unless postmedieval-era sources contained new, reinterpreted, or heavily re-written evidence, duplications were not included in HRDL. Only a single example was required for HRDL and analysis, to avoid duplications dominating and therefore biasing the corpus.

Archaeological evidence from 20th and 21st century cathedral excavations revealed 'medieval' and 'postmedieval' evidence of historic burial management, although not always dateable to the decade or century. Thus, 'medieval' and 'postmedieval' terminology applied by excavators for relative dating of evidence, based on PAS periods, could easily be incorporated into HRDL.

2.2 Finding sources

Prior to this project there were no primary or secondary sources listing examples of exhumations for any of the thematic groupings. Starting from scratch, divergent methods were needed to find medieval versus postmedieval texts, further underpinning the medieval/postmedieval division.

2.2.1 Handlists.

If a sub-corpus could be built from footnotes, bibliographies, and indexes in secondary sources, researchers created a 'handlist' in Excel for that sub-corpus which summarised the texts' relevance and where the base copy could be sourced. Short descriptive summaries – limited to two sentences were added to each – written by the researcher responsible for their collection, for two reasons: 1) such information varied considerably between sources from whence material was collected, 2) to avoid potential copyright infringement. Bibliographic details were captured in the handlist and Zotero to avoid duplicating entries. These handlists were converted to .csv (comma separated value) format to provide a simple means of

Just Accepted

populating the final digital library build that could be easily managed for future maintenance and updating, without requiring knowledge of SQL database functionality or management.

Medieval handlists leaned on existing scholarship where saints or royals had been exhumed or specific medieval authors were known for mentioning exhumations. Targeted searching for individuals subjected to exhumation or known to discuss it, actually produced a broad spectrum of medieval languages and genres, from the biographical and hagiographical to legal and theological treatises, providing an additional level of categorisation.

Since saints were commonly exhumed before the Reformation, the 'Saints' handlist began with published lists of medieval saints in Britain and their primary sources [12–14]; recent scholarship of well-documented saints, medieval authors commentating on mortuary practices and general cult of saints scholarship [15–18]; and digital databases of saints.

Archaeology began with the few published cathedral excavation reports, then focussed on the many unpublished reports ('grey literature') available from the Archaeological Data Service (ADS) and guided by the Historic England NRHE Excavation Index for England (EI).

2.2.2 Online archives.

These handlists were then supplemented by searching online archives; a method used for the other sub-corpora as well. The sheer volume of material available for each sub-corpora meant finding and filtering relevant sources required as many methods and regular expressions as there were languages, sub-corpora, and relevant digital archives [19]. This approach produced an unviable amount of material from multiple periods and genres. Combining traditional archival work to find and manually digitise specific accounts with wide-spread searches of multiple online archives meant sources known and unknown to the team were captured, and while not exhaustive, it is certainly a substantial and unprecedented corpus of exhumation material.

The term 'exhumation' (and its variants) was rarely used in English until the late 1700s [20]. Terms such as 'disinter', 'unearth', 'uncovered' or 'dig/dug up' tended to be used instead, but even these words were insufficient for finding relevant accounts of exhumation since authors might only refer to specific human remains (e.g. skull, ribs, remains); use archaic or poetic terms (e.g. ashes, sepulture, entomb); only mention the location of the remains (e.g. crypt, vault, tomb); or focus on 'reburial' and its synonyms. Thus, sub-corpora often needed their own shortlist of keyword search terms, which varied by language and period, since spellings and words evolved. For example, the term 'charnel' was inadequate by itself, so Farrow developed a corresponding keyword vocabulary for identifying charnel through adjacent terms (e.g. 'skulls', 'bones', 'remnants'). Keywords were later expanded to incorporate specific site/church name(s) sometimes in addition to charnel-related church features such as 'bone-chutes' and 'bonehouse' (each with variant spellings) identified during research. This yielded a higher proportion of false positives as the terms were broader but was essential for identifying site-specific examples.

2.3 Greps and Regexes

More sophisticated user-interfaces for certain online archives were far more accurate and efficient for finding relevant material. *The Dictionary of Old English Online Corpus* (DOE Online), the *Consolidated Library of Anglo-Saxon Poetry* (CLASP) database, and *The University of Michigan Library* interfaces for collections such as Corpus of Middle English Prose and Verse, Early English Books Online (EEBO), and Eighteenth Century Collections Online (ECCO) produce concordance-led search results which aided discovery. This facilitated rapid filtering and search refinements within the user interface, rather than downloading texts and searching them as a separate step.

For digital archives with simpler search facilities, such as Internet Archive, the keywords usually produced numerous results. The effectiveness of keyword searches for surfacing relevant accounts was explored for the 'cathedrals' sub-corpora as it was the largest single collection identified through Internet Archive prior to sifting out irrelevant records. As Table 2 below shows, of the 536 files uncovered on Internet Archive at the time of search using the identified Early and Modern English keywords (using seven

Just Accepted

cathedrals as a sample), only 3% (18 records) contained relevant material for HRDL. This partly reflects 20% being duplicate files, but the remaining 77% irrelevant files either used keywords in a different context (e.g. poetically or metaphorically); were only a partial match (e.g. 'bury' in 'Canterbury') or had multiple meanings (e.g. 'translation' can mean a saint's exhumation or translating a language). Thus, keyword searching alone was unreliable, and results varied per archive based on the quality and clarity of the metadata and scanned material. This justified our re-OCRing of material, improving the accuracy and efficiency of results.

Table 2: Results of search parameters for seven cathedrals on Internet Archive

Cathedral	Documents	Used	Unused	Duplicates
St Giles	20	4	15	1
Glasgow	23	7	13	3
Chichester	60	2	39	19
Durham	178	1	144	33
Ely	98	1	76	21
Hereford	114	1	87	26
Manchester	43	2	37	4
Totals	536	18	411	107
% of Total Found	(100%)	3%	77%	20%

Even within a relevant text not everything mentioned met HRDL's thematic remit. For example, medieval chronicles might include references to exhumations of royals and saints, but also list other historical events (wars, famines, politics etc.) not relevant for HRDL. Similarly, a 19th century antiquarian's gazetteer of churches they visited could include tombs they opened but also the church's history and architecture, which was superfluous for HRDL. Developing a procedure for quickly and consistently combing large amounts of collated material was thus imperative.

Rather than manually reviewing thousands of potential texts for relevant passages, the team used Grep software (*AstroGrep* [Windows], *DocFetcher* [Mac]). First, relevant texts were downloaded and converted into plain text (.txt) to make them readable on all project software. Once the Grep software had compiled all texts into a viewfinder, bespoke regular expressions (regex) [19] were used to search the material (apart from Welsh: see below) and highlight matches across every text. This made it faster and easier to scan for relevant content through immediate context.

Given the level of standardisation and evolutions in spelling forms throughout the twelve centuries and multiple languages covered by the project, broad and flexible regex queries were required to effectively scan sources. To achieve this, the team compiled lists of words that might refer to burial, exhumation, human remains, or mortuary practice for each language. The Old and Middle English lists drew on the *Thesaurus of Old English* [21], Middle English Dictionary, and CLASP's *Word-Hord: A Lexicon of Old English Verse* [22], while the Latin collation used the *Dictionary of Medieval Latin from British Sources*. Each list was then ordered into a single string and reduced to their base common elements with 'wild-card character' substitutions for suffixes and prefixes. For example, *rebury*, *bury*, *buried*, *burying*, *burial* etc. could be reduced to the search formula: *buri*|*bury*. Since most spelling variations revolve around vowels, such as vowel mutations or regional stresses, most vowel positions were turned into wild card characters, especially for medieval language searches.

As greps highlight words within their sentences, we could quickly identify relevant from irrelevant words using the regex. Passages pertaining to exhumation often appeared in clusters, with multiple words highlighted within the same paragraph, e.g. an account would often mention 'digging', 'coffin', 'shroud', 'tomb' in quick succession. This was particularly important when working with Welsh material.

2.4 Challenges with Welsh-language texts

Variations in in Welsh word forms are very different to those in English, with the grammatical systems underpinning initial phoneme mutation, noun gender, and pluralisation mean that Welsh words regularly

Just Accepted

vary in distinct manners. The inconsistency in medieval variation further increases the potential variance observable in cross-period forms.

This varietal range meant that searching online repositories by only using a word's standard modern spelling, namely the radical (non-mutated) masculine singular form, would return only a small fraction of relevant results. As such, rules for spelling variations caused by medieval inconsistencies and initial mutation in both medieval and modern forms were defined to capture all possible forms of each keyword [23]. It was decided that creating rules to account for the various ways of forming the plural from a root stem would result in too many non-existent forms, and so both singular and plural forms of nouns were included separately on the base keyword list.

From 42 initial keywords related to exhumation,12,228 potential Welsh spelling variations were determined from the preliminary list of potential variables; e.g. *datgladdu* (to exhume) alone yielded 3,564 possible permutations. Some of these forms were highly unlikely to have ever been used, e.g. *nadcladdu*.

The amount of hits returned from this initial generation was far too large a number for manual management (with results exceeding the hundreds of thousands), and so instead of creating regex to create a wider linguistic net, the Welsh search terms and spelling variables were reduced and refined in order to reduce the number of irrelevant results. This included the removal of all phrasal verbs, e.g. verbs calquing on 'digging up' (*palu lan*, *palu i fyny*, *palu y fyny* and their respective variants), which were particularly problematic as they included the immensely common words *i* (the preposition 'to' and the pronoun 'me/my') and *y* (the definite article 'the', and, in Middle Welsh, the preposition 'to' and pronoun 'me/my'). The term *arch* (coffin) was removed for the same reason as it can also mean 'command', 'prayer', 'chest', 'torso', '(Noah's) ark', and it appears in many compound words, e.g. *archesgob* (archbishop), which generated numerous irrelevant results. The dash in words such as *ail-agor* (re-open) were also deleted as this returned far too many instances of *ail* on its own, which means 'second'.

Despite this refinement, EBBO and Welsh Newspapers were still returning too many irrelevant results, so the search parameters were further reduced to only the forms and variants involving *datgladdu* (to exhume), *corff* (body), and *dadorchuddio* (to uncover). Specific keywords in context reviews were undertaken to identify suitable material, for example, *corff* (body) highlighted in close proximity with any verbs involving 'moving' or 'exhuming' was indicative of a relevant passage. More often than not, if a passage had *corff* and no other words highlighted, it could be instantly ignored. Using semantic context also reduced confusion where potential permutations intersected other word forms, such as seen with *medd*|*med* (generated from *bedd*, 'grave') and *medd* ('mead').

2.5 OCR issues

However, the grep-stage of the workflow indicated that poor-quality document scans were producing plain text files so 'dirty' they were unreadable by greps, or even by human eye. Therefore, rather than trying to decipher the dirty TXT file or manually clean potentially relevant sections, we re-OCRed every scanned document first, except archaeological reports which were not being converted into verbatim extracts (see below) as a consequence of the full reports being published works under copyright – each record is referenced for a user to source the original material. A bespoke OCR process was created that utilised the Google Vision API (the best 'out-of-the-box' tool for character recognition for 60+ languages) alongside Tesseract (which outperforms Google Vision for layout detection but is poorer for character recognition). This new method was highly effective in improving the quality of the PDF–TXT conversions and has been published separately [24].

Despite the overall accuracy in transcribing words without any degree of autocorrection, a number of semi-consistent OCR mistakes did remain which are common in historical texts. These could not always be corrected by the re-OCR process, so light manual cleaning was conducted once extracts were ready for HRDL. However, even the most muddled of OCRed texts could be quickly cleaned following re-OCRing. The occasional in-text citation in another language — quotations of Greek and Hebrew in antiquarian sources (18th and 19th centuries), for example — also had to be identified and translated. Where an extract

Just Accepted

switched to or quoted another language without translation, the original passage was retained and a translation in modern English was provided in brackets for the medieval sub-corpus.

2.6 Challenges with medieval texts inc. duplicate saints/royals

The project did not use handwritten documents but rather typed material. The *DOE Online Corpus* [25] and *CLASP* databases provided cleaned digital texts which facilitated grep searches in Old English, but the Middle English sub-corpus proved more problematic. Although many clean Middle English texts are available online via *Corpus of Middle English Prose and Verse* [*26*] other texts were only available in print volumes with archaic typefaces. Google Vision API and Tesseract both had difficulty recognising runic and medieval characters frequently used in Middle English — like ꝑ, þ, ð, ȝ — and offered muddled transcriptions. Moreover, medieval English contains many vowel mutations, as well as regional variations and scribal interpretations of how a single word can be rendered, similar to the Welsh corpus. For instance, the Middle English word *grave* (grave; burial place) is also found as *grawe*, *grafe*, *greve,* and in various declined and poetic forms. To overcome this, we made the regex string broad enough to accommodate potential spelling attestations and OCR errors (see Middle English entry in [19]). We also used *FolioViews* to analyse the Old English corpus since it refined our regex strings and helped identify attested spelling variations or permutations.

2.7 Embargoes

Farrow's 'Charnel' sub-corpus, as the basis of his doctoral thesis, had a significant portion or records embargoed from HRDL until after his examination and thesis publication. Late medieval charnel extracts were contextual and therefore suitable for earlier release. A small number of other extracts supporting active research and publication efforts were similarly embargoed, but all records are live as of publication.

2.8 Challenges with archaeological reports

The 'Archaeology' sub-corpus contains 174 records of 20th and 21st century burial excavations from 42 British cathedrals, at launch: the most comprehensive survey of its type. Grey literature presented a number of discrete challenges, as not all reports were available as digital downloads, with both physical and digital access being inconsistent across archives or current content holders. Authors, commercial archaeology units, Historic Environment Records (HER) and cathedral archives were usually able to supply the data directly, but only upon request and for material specified. A published, updated list of Cathedral archaeologists and their respective work(s) could expedite the discovery and supply of such data in future.

Excavations could also reveal interventions from across multiple periods based on the reported evidence. Thus, if a grave had been robbed in the 17th century and later truncated by heating pipes in the 19th century, this would provide the basis for two separate events – and two corresponding records to be created, grouped under the HRDL's 'Related Records' structure. Unlike historical documents which were atomised into individual extracts, the archaeology sub-corpora comprised .txt files containing a summary of the 'event' by Daubney with a consistent set of details summarised from the broader report, including aspects relating to the treatment of human remains, locational specifics, grave goods and furnishings, and subsequent actions taken. Given this level of abstraction from the original reports, these records were kept separate from the other full-text material through a discrete metadata field that captured their Evidence Type as 'Archaeological'.

The same metadata fields used for other sub-corpora were applied at the document-level, meaning archaeology was fully aligned with the historical documents. Forum on Information Standards in Heritage (FISH) vocabularies [27], an industry-standard in UK cultural heritage, were used to record types of archaeological items, making HRDL searches logical, consistent, and intuitive for archaeologists. However, motives for burial management and the treatment of human remains or artefacts during and after excavation required translating into standardised vocabularies [28].

Just Accepted

The terminology in archaeological grey literature was highly variable, e.g. 'charnel', could be used to describe both intentional deposits of comingled human remains and unintentional spreads of disarticulated human bone (see 'Challenges with Postmedieval Sources' above). Similarly, where/whether human remains or grave furnishings were to be stored or reinterred afterwards was inconsistently reported, whereas location information and notes on other finds was consistently detailed. Aligning excavations with historic texts was key to our thematic dataset.

Archaeological investigation in cathedrals aims to minimise disturbance, restricting the scope of excavation accordingly, and only reporting intentionally uncovered archaeology – not all the archaeology actually present. This restricts interpretations of the spatial and chronological trends in historic burial management and required a synthesis of material to capture aspects per (historic) event rather than per (modern) excavation to meet HRDL's thematic focus.

3 EDITORIAL DECISIONS

3.1 Footnotes

Footnotes were common in many types of documents, and sometimes contained (the only) information relevant to the library. Since .txt files do not preserve superscript numbering systems for footnotes, they could cause confusion for HRDL end-users. However, as footnotes can represent different authorial or editorial voices (e.g. authors' commentary on a reprinted work) this poses a problem when developing a digital library built around chronology. To compromise, footnotes with important standalone information were converted into a separate extract, while footnotes with contextual commentary were retained as separate files for research but not for HRDL. Minor relevant footnotes were treated as standalone text within the same extract; irrelevant footnotes were deleted.

3.2 Translations

Multiple languages needed to be navigated: Latin, Old English, Old Norse, Anglo-Norman French, Middle English, and Medieval and Modern Welsh. Rarely do digital libraries of this scope and scale cover so many historical languages at once and the team had to be diversified to identify and translate sources from these languages. However, we identified early on that a single regularised language (Present Day British English) would be required by stakeholders, simplifying the search syntax for end-users who did not have a specialist background in historic languages, and for dedicated analysis to be able to be undertaken from a common linguistic base.

Multilinguistic analyses are very difficult and time consuming to achieve with Big Data corpora, and there is negligible valuable in the extra efforts and specialisations required for conducting cross-language analysis: instead, translating multiple languages (or diachronic/unstable variants) into a common form has produced viable results of comparable significance on other projects where emphasis is on linguistic quality and intra-content comparison [29,30].

If the passage was not in Modern English, and no existing translation could be sourced, a translation was produced in-house by the team as an extract. Our metadata model allowed us to record what the original language was for reference, and the original languages would be available to download from a separate page of HRDL for use by researchers. That way, end users can engage with sources in original or modern languages.

3.3 Cleaning and Spelling Normalisation

Texts were first cleaned through a bespoke Python script written to correct line-breaks, hyphens, numeric abbreviations, errant spacing, reference symbology, and other common OCR issues with carefully defined find-and-replace functionality [31]. *VARD* was then used to normalise early modern and non-British English spellings across the corpus to modern British English [32], with a 65% automatic normalisation threshold and manual approval for remaining tokens. This process replaced American English, Australian

Just Accepted

English, and antiquated British English spellings, since the content of a text was prioritised over specific linguistic forms.

4 WORKFLOW AND EDITORIAL DECISIONS

Having outlined the general workflow for the digital library's texts, we now detail the challenges encountered in developing specific sub-corpora for reproducibility and transparency, in the hope of supporting others seeking to work with these types of genres or period-specific documents.

Once sourced and prepared, the original document and its resultant .txt files ('extracts' corrected to our standards) were stored together in a 'Family Folder'. The 'Family Folder' system used a standardised file naming hierarchy which was critical for the HRDL's overall structure (Fig 1).

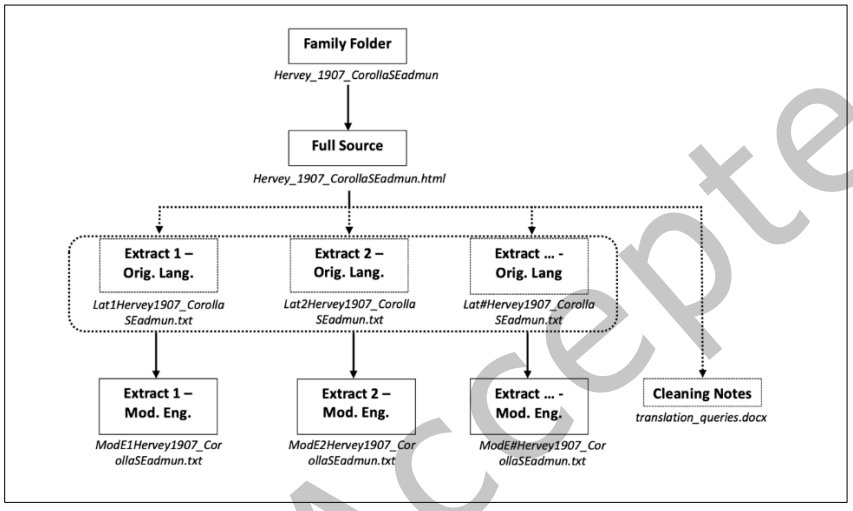

Figure 1: HRDL Project File naming hierarchy

Each extract file shared the same filename as the Family Folder but was prefixed with (a) the language of that file (e.g. 'ModE' for Modern English or 'Lat' for Latin) and (b) a number, to differentiate extracts and language-state from the same source.

Numbering started at '1' for each new language group in a Family Folder, which was important for multilingual source material. Extracts were numbered in the order in which they were found within the full text. Applying the same naming convention to the folders and their related files preserved the connection between files derived from a single source. It also meant we could track the progress of a Family Folder and its contents as a unit throughout the workflow so that all related extracts could be quickly identified at any stage in case an issue arose. This relationship appears as 'Related Records' in HRDL to assist intra-textual analysis.

Each extract was checked at this stage, correcting any errors introduced or missed by the OCR process, adding any missing vital information to the extract in square brackets (such as the year or name of the deceased) mentioned earlier in the original text but absent from the individual record; and removing any footnote numerals or critical apparatus missed by the automated cleaning process outlined above.

4.1 Metadata

Metadata for each extract was simultaneously recorded using a live Excel document per sub-corpus, termed 'Trackers'. These allowed us to track the progress of each sub-corpora and any missing or ill-defined metadata. One researcher was assigned per sub-corpus and tracker to maintain consistency. As discussed below, metadata fields were based on Dublin Core while also mapping to CIDOC-CRM principals

Just Accepted

(see [28]). We standardised naming conventions for each of those metadata fields, such as the listing of individual persons and their titles.

To manage the variety of texts encountered, an additional field was added to the core document-level metadata to capture Reference Type, with three values used to group texts: (1) Passing Reference (2) Anecdotal (3) Descriptive Narrative. A 'Passing Reference' is an extract with very little information such as a single line entry in a liturgical calendar or a brief footnote, without extensive detail or description. These were the least impactful for understanding the wider semantic questions of the project and stakeholders, but provided invaluable evidence of specific events to help create a timeline and were useful on aggregate to demonstrate repeated concepts over time.

An 'Anecdotal' extract presents more information than a 'Passing Reference' (e.g. 2–3 paragraphs), concerning the nature and cause of the event and the people involved. These include extracts from hagiographies, compendiums, gazetteers, and early guidebooks etc. 'Anecdotes' are useful for both close and distant reading of the corpus. 'Descriptive Narratives' was reserved for lengthy extracts which did not need to be subdivided into separate extracts. These include pamphlets, essays, reports, and newspaper articles, often containing emotional responses, theological or ethical concerns, wider debates, and highly detailed explanations of exactly what was found. This additional metadata field demonstrates the value and importance of maintaining a carefully controlled vocabulary, in order to ensure internal consistency whilst allowing sufficient flexibility to facilitate nuanced (re)searching.

The potential values for each metadata field were identified by thoroughly refining potential free-text options identified while building the sub-corpora. Additional expertise from project advisors guided the choice of specific terminology to provide a vocabulary that bridged the academic domains most likely to engage with HRDL. Developing a unified controlled vocabulary for archaeology has been identified as a discipline-specific priority [33], which helped guide our definition of potential values for a series of additional fields designed to capture and regularise contextual metadata from the extracts as a means of aligning with and bridging adjacent discourse.

We originally approached Zooniverse for crowdsourcing this contextual-level metadata, however their preference for free-text entries to make data capture more engaging for their users went against our need for a controlled vocabulary, and as a consequence we felt this would not be an appropriate platform for capturing additional metadata.

4.2 CIDOC-CRM for Metadata

HRDL required a core underpinning metadata layer that encompassed both standard Dublin Core fields as well as additional aspects that could be codified without relying on specific keywords being present within an extract. This metadata was built around a bespoke ontology developed using the CIDOC-CRM framework; the value and challenges of this cross-framework adaptation are outlined in this section.

CIDOC provides a Conceptual Reference Model (CRM); an adaptable ontological framework to standardise the interrelated values of, and between, any given entity/ies for building a relational hierarchy. CIDOC-CRM was developed primarily for the controlled exchange of cultural heritage and museum documentation between their collections. It comprises a rigidly structured formal hierarchy of concept class and working relation property between these values. The tightly controlled structure was primarily designed for physical collections and entities, particularly archaeology [34,35], but the versatility and ongoing development of the ontology has allowed for its integration across a wider range of digital resources. The CIDOC-CRM is known to be viable for modelling the bibliographic content of library collections [36,37]. The Dublin Core ontology for bibliographic metadata has previously been mapped onto CIDOC-CRM [38,39] which provided much of our metadata at a document level. Yet given that HRDL's metadata must also capture distinct thematic content (i.e. exhumation evidence), these aspects required a bespoke framework that expanded the core Dublin Core content and intersected it with our custom metadata values, such as the 'Account Type' delineation outlined above.

This bespoke mapping involved referring to other published models (including OpenAtlas), to ensure HRDL's ontology had an interoperable basis, and specifically aligned with similar (meta)data mapping by

Just Accepted

a CRM data model for early medieval cemeteries in Austria by The Anthropological and Archaeological Database of Sepultures (THANADOS) project (2019–21) [40]. The THANADOS model primarily concerned archaeology (graves, individual burials, finds — see Figure 1 in [41]) encompassing many of the fields we needed to capture. Combining Dublin Core metadata with a CIDOC-CRM ontology therefore meant the thematic categorisation of HRDL's corpus could transcend and expand the boundaries between historical texts and archaeological data.

4.3 Evaluation of Platforms

Many major digital libraries have been built around TEI-encoded texts (e.g. [42]), and although this was initially considered, the processing workflow we designed meant annotated extracts were superfluous. Textual content was the capture priority rather than information relating to original physical formats, which were rendered moot by our extract creation process, and as all our records were translated into a common language with consistent spelling throughout the collection, there was no need for such internal annotation. Furthermore, serving primarily as a knowledge-exchange digital library, our end-users were not concerned with running TEI-based analyses.

HRDL is not aimed at digitising and interpreting physical manuscript pages; indeed, many of the texts compiled for HRDL were too poorly printed or used complicated fonts that were barely readable in their original form, so presenting the original manuscripts was counterintuitive to making the information easily accessible for a non-academic audience. Therefore, digital library models based on viewing and interacting with PDFs or images of historical manuscripts (e.g. *Internet Archive*; *Roman de la Rose/Christine de Pizan Digital Scriptorium; The British Library manuscripts collection*; *Eighteenth Century Collections Online)* were not useful for HRDL. As HRDL collates digitised (and often OCRed) texts, the underlying intention is not solely the digital preservation of specific manuscripts for an academic audience but creating a thematic collection of texts for active (re-)use by practitioners to inform planning, policy, and public engagement.

Both Omeka and CKAN were initially explored as platforms for HRDL. Originally designed to support digital collections of objects and images, Omeka can be adapted for visualising textual information. However, it had limited options for formatting and displaying plain text, especially if they were not TEI-encoded. Omeka relied too heavily on third-party plug-ins to support features HRDL would need, and if these plug-ins were discontinued or not updated, the library would lose functionality. Similarly CKAN, an open-source data management system which has been used for digital humanities collections [43] could be adapted for a digital library. Being open-source enhanced sustainability without relying on third-party plug-ins. However, CKAN required a steeper learning curve to adapt it for use as a digital library in terms of the search facility and the visualisation of the text files since it is used more for large numeric datasets. It was essential the UI was accessible for those with no experience of navigating academic databases, and we required display flexibility so search results could be visualised with multiple elements of the text and metadata appearing initially for each item returned by a search (title, description, document ID), and then for each record in full; the display of both the search returns and the individual records was paramount. This two-step search model was not supported by Omeka and too complex for CKAN.

Additionally, two separate search boxes for the 'names of the deceased' and the 'people involved' were also incorporated, providing auto-complete prompts as specific name form and formatting was identified as a significant barrier to finding material related to desired individual(s). Ultimately, we needed a dynamic search engine to enhance the user-friendliness and accuracy of the library, enabling content discovery from multiple angles.

5 DESIGNING THE LIBRARY

The design philosophies of HRDL's user interface and search facility developed by Butler with the University of Liverpool's IT Services team, was structured around non-specialist user accessibility at every step. This included four core aspects: (1) The development of a codificational strategy in order to control vocabularies and ensure related content was consistently grouped; (2) randomised search results as the default display option on HRDL's landing page to encourage further exploration of records beyond the

Just Accepted

immediate interests of the end user; (3) the combination of non-specific free text searching, with static filtering provided through a range of metadata categories; (4) the manner in which search results are displayed, as tiles with a short but catchy descriptor.

In light of the limitations of the existing platforms evaluated and outlined above, a bespoke digital library was required to meet the project's needs. HRDL was built using the Blazor framework which was chosen for four key reasons: (1) increased performance and security, particularly for the protection of the host institutional servers; (2) reduced complexity for the integration and maintenance of front-end tooling, designed specifically for the use of C#; (3) form value retention, allowing users to navigate pages and full records using browser buttons or commands without resetting filter options; and (4) enhanced cross-platform development, enabling HRDL to work on devices with different screen sizes. The digital library served as a pilot project for the University of Liverpool's IT development team to explore this relatively recent framework, which received institutional approval and its success with HRDL supports the case for being adopted for future projects.

The key principle underlying HRDL's interface design was an enriched, flexible search facility, allowing detailed filter options whilst presenting records in a manner likely to capture attention, providing potential angles for further investigation, and ultimately encouraging exploration of the material. We sought to simulate the browsing of a library's bookshelf, offering related material that might be found alongside a desired record: a core stratagem for increased engagement [44] that focusses on increasing the volume of 'areas of interest'. Our creative decision to randomise the display of records by default (with options to re-order by publication date or arbitrary document id), in order to showcase the range and representativeness of material available in HRDL. This design followed a core tenet of linguistic concordance analysis, where random ordering is essential to minimise solo document bias and provide a broader overview of the desired search term(s). Exposure and positional bias has a recognised impact on all aspects of digital resource design but 'recommending results' has received marked attention as a useful component for facilitating extended use [45,46], and particularly in encouraging end users to explore new lines of inquiry.

Metadata filtering had to be accessible at every step of user interaction, allowing dynamic refinement of search results as new information was identified by users but without resetting filters when returning to the base search results. Users can preview each record and its core details (full description, full title, body text) then the full record: a key strength of Blazor [47]. The relevant metadata values are provided, each allowing additional record filtering of the base search results that led to that record.

Given the importance of named (and unnamed) individuals noted throughout the records, both as the subject of discussion and as participants, two design considerations were implemented to surface this information. In addition to the core keyword search provision, two additional free-text search bars that operate only within their respective metadata fields ('Name(s) of Deceased' and 'Person(s) Involved'), were added as additional dedicated filter options.

6 DIGITAL LIBRARIES FOR MORE THAN RESEARCH: THE HUMAN REMAINS DIGITAL LIBRARY IN CONTEXT

The immediate digital landscape for resources closest in scope and focus to our own demonstrates why our bespoke, adaptive approach was necessary. With a dataset bridging church histories and individual physical archives [48], archaeological reports (see [49–51]) and cultural engagement [52], there was no existing model that encompasses and bridges all these relevant aspects. All the digital resources featured in the studies cited above were designed to meet the researcher's own needs and designed to address specific research questions or provide new specialist datasets. THANADOS [40] exemplifies digitised archaeologically-led resources, where interactive elements such as charts, highly specialised searches, and base GIS services enriched with finds and site details, all offer new means of browsing standard archive data. Resources created through and for knowledge exchange, such as HRDL, can generate new engagements or collaborations with non-academic audiences and new mediations that challenge traditional research or sector distinctions and dynamics, but it requires collaborative, interdisciplinary

Just Accepted

innovations tailored to both support and transcend these boundaries and expectations. As outlined above, our digital library illustrates how this can be done through building on enrichment opportunities whilst retaining a core ontological framework [53], to directly address the unique requirements of our cross-sector stakeholders.

The Human Remains Digital Library was designed with a broad user-base in mind, spanning an array of academic and technical experience; catering to both sectors equally while ensuring accessibility and user-friendliness for non-specialist audiences. Our end-users may need to explore British reburial practices in a meaningful manner whilst simultaneously conducting specialist analysis in a research-driven context. Where we differ notably from a "typical" digital library (if such a concept can exist outside of standard expectations for such resources) is that we do not prioritise the display and composition of original manuscripts, sources, or forms; in translating multilingual records to Present Day British English; in not having a primarily linguistic focus; and in combining archaeological and textual evidence. As such, information quality and fidelity are key priorities for HRDL, given its singular thematic focus.

The underpinning Big Data model guiding our design is not just modelling the how and where of potential physical human remains evidenced from the accounts that comprise our digital library, but on further engaging with the 'eco-dynamics of digital cultural heritage' [54] whereby the broader beliefs, protocols behind, ideas of, and concepts of respect relating to them can be engaged with through the manner of reporting. Bringing together both sides of analysis, what is said and what is signified, lies at the very heart of discourse analysis; it is only though creative efforts to bring these two aspects into closer alignment — through such comprehensive metadatic and semi-standardisation as outlined herein — can the nuanced relationship between specific form, social context, and intended meaning be more fully captured, and thematic digital libraries capture meaningful hermeneutics beyond just the linguistic.

7 CONCLUSION AND RECOMMENDATIONS

7.1 Thematic Digital Libraries for Knowledge Exchange

HRDL shows that a thematic, multilingual, cross-evidential library can be engineered to meet both research demands and the operational needs of living heritage institutions. Thematic digital libraries, where content is scattered across many repositories, evidence types, genres, periods, languages, and authors, require a level of standardisation different from digital libraries built for manuscript analysis. HRDL's workflow required a combination of existing knowledge and discovery-led searching for sources, supported by handlists for known examples; multilingual/variant spelling keywords developed from knowledge of poetic, archaic, and evolving word usage for discovery-led searches; and concordance-led greps or to rapidly identify relevant passages within large collections.

For any Big Data digital library, the re-OCR of legacy scans can substantially reduce editorial workload and improve text accuracy. Where cleaning may still be required, bespoke Python scripts can correct predictable issues (e.g. line breaks, floating commas), and strip-out unwanted elements (e.g. footnotes). Whilst digital text correction toolsets can provide significant benefit in terms of reducing editorial resource outlay, they can also introduce additional errors; particularly with regards to antiquated language, as found in the majority of the material which comprised our various sub-corpora and thus required additional preparatory steps.

7.2 Historical Multilingual Corpora and Digital Libraries

Using online dictionaries for historical languages to compile search terms should include both specific and general terminology whilst considering carefully possible spelling variation (whether grammatical, intentional, or accidental), and multi-period corpora should incorporate both modern and historic terminology. For multilingual corpora of historic languages, modern translations and standardised spellings make corpora more accessible and searchable: a valuable priority for knowledge exchange audiences, with original language/spelling corpora available for separate bulk download to satisfy academic audiences too. These separate files can be easily annotated and/or analysed by end-users

Just Accepted

without needing to strip out existing schema. Similarly, tools such as *VARD* offer rapid spelling normalisation and the effort of refining the underlying dictionary can be worth it for parsing large thematic collections which may likely house unusual domain-specific and historical terms and spellings in greater numbers than non-thematic corpora. Beyond providing guidance, our implementation contributes a tested pattern for cross-language normalization and excerpt curation that preserves scholarly fidelity while maximizing public usability—an approach readily transferable to other domain-specific corpora.

7.3 Combining types of Cultural Heritage

Combining archaeological evidence with historical sources is currently rare in digital libraries. HRDL enriches Dublin Core with content-level contextual metadata, mapped to CIDOC-CRM, to align heterogeneous textual and archaeological evidence without TEI encumbrance. This meant combining controlled vocabularies from both The Library of Congress (historical texts) and FISH (archaeology) while also standardising titles, names, placenames, actions, and motives not covered by the controlled vocabularies ensured intuitive searching for both evidence types (and end-users) while maintaining consistency across, and visibility of, all metadata and texts. This integration enables queries that cut across evidence types (e.g., event motives, treatment of remains, spatial context) and supports decision-making in conservation, planning, and public engagement.

Our Blazor implementation provides a responsive, secure, and accessible interface with form-retention and dynamic filtering tailored to non-specialists, validating the framework for future heritage resources. Underpinning this with a relational hierarchy based on an interoperable ontology for bespoke cultural heritage collections (e.g. CIDOC-CRM) facilitated the integration of different entities and evidentiary types in a way that transcended the boundaries between the sub-corpora, creating a unified thematic corpus. Current out-of-the-box platforms for digital libraries, especially those relying on third-party plugins for required functionality, are not yet able to support this. Moreover, learning CIDOC from its case studies and other users (e.g. THANADOS) meant HRDL benefitted from other academic and non-academic knowledge exchange as well. Citizen science platforms (e.g. Zooniverse) offer further knowledge-exchange opportunities on a global scale. However, in seeking to collect (meta)data through such platforms, carefully consider the types of tasks volunteers will engage with and whether they, and the project's topic, is suitable for the 'gamification' ethos of some platforms. Also anticipate that some platforms may seek to direct the research as a stakeholder rather than act as a service. Taken together, HRDL provides a reusable blueprint for thematic digital libraries that must bridge sectors, languages, and evidence classes. It transforms dispersed historical and archaeological records into an actionable corpus for practice, and a scalable dataset for research.

Thematic corpora are under-represented in the digital library landscape, yet they are pivotal for engaging non-academic audiences on sensitive topics such as exhumation. Maximising skilled search-and-recover methods across multiple repositories to form a unified corpus of evidence, and crafting multi-layered search functions in the UI to support both strategic and discovery-led queries means end-users' need for answers to important ethical, practical, theological, and historical questions — from conceptual to site- or date-specific — are at their fingertips, fully embodying the concept of reutilisation and the generation of new knowledge from old sources brought together under a new context [55]. Refining, regularising, and working to controlled vocabularies that meet the needs of different end-users — whether stakeholders, anticipated audiences, or academic practitioners — during the build is vital for enhancing the utility of a digital library, in order to ensure longer-term engagement and legacy beyond just the research priorities underpinning their creation, which remains one of the most significant hurdles for such resources. By consolidating 'old' sources under a new, interoperable context, HRDL demonstrates how knowledge exchange and scholarly inquiry can be advanced simultaneously, offering a sustainable model for future heritage-focused digital infrastructures.

Just Accepted

ACKNOWLEDGEMENTS

The authors would like to extend their strongest thanks to prior members of *The Human Remains Project*: Dr Niamh Kehoe; Rosie Djurovic, and Dr Rachel Swallow; and to the project residents: Dr Alexandra Claridge; Dr Olivia Colquitt; Karen Morrell; Dr Richard Bellis, Dr Brian Costello, and Dr John Jenkins for their irreplaceable contributions. Further thanks go to Dr Tom Lynch (The British Library) for help with sourcing texts during UK Covid-19 lockdowns and to the CLASP project for early access to their data. We are grateful to the volunteers who tested Zooniverse workflows; members of The Human Remains Collaboratory for vital feedback and beta-testing; to Emily Lanigan-Palotai (Chester Cathedral, England), Tom Livingstone (Heritage Stonemason), and Fran Hollinrake (St Magnus Cathedral, Scotland) for contributing archival texts. Assistance provided by Darren Hoy from University of Liverpool's IT Services is also greatly appreciated.

The authors would like to acknowledge and thank the funding and support of the UKRI Future Leader's Fellowship 'The Human Remains: a digital library of British science and mortuary investigation' [MR/S032800/1 and MR/X023567/1], at the University of Liverpool, UK.

Just Accepted

DATA ACCESS STATEMENT

Data supporting this study are openly available from HumanRemainsDL (GitHub) at https://github.com/HumanRemainsDL/.

OPEN ACCESS STATEMENT

For the purpose of open access, the authors have applied a Creative Commons attribution (CC BY) license to any Author Accepted Manuscript version arising.

REFERENCES

[1] Liv Nilsson Stutz. 2023. Between objects of science and lived lives. The legal liminality of old human remains in museums and research. *International Journal of Heritage Studies* **29**, 14 pages. https://doi.org/10.1080/13527258.2023.2234350

[2] Sarah Tarlow. 2023. The Names of the Dead: Identity, Privacy and the Ethics of Anonymity in Exhibiting the Dead Body. *Public Archaeology* **0**, 20 pages. https://doi.org/10.1080/14655187.2023.2268384

[3] Ruth Nugent. 2020. Burials in Churches. In *The Building Conservation Directory Special Report on Historic Churches: twenty-seventh annual edition*. Jonathan Taylor (Ed.), Cathedral Communications, 4 pages.

[4] Ksenija Bogetic, Vuk Batanović, and Nikola Ljubešić. 2022. Corpus compilation for digital humanities in lower– resourced languages: A practical look at compiling thematic digital media corpora in Serbian, Croatian and Slovenian. *Suvremena lingvistika* **48**, 23 pages. https://doi.org/10.22210/suvlin.2022.094.01

[5] Ruth Nugent. 2016. *Handling the dead: a haptic archaeology of the English Cathedral dead*. Ph.D. Dissertation. University of Chester, Chester. http://hdl.handle.net/10034/620390

[6] Brian Costello and John Jenkins. 2023. The Hospitaller preceptory of Slebech, Pembrokeshire: Interpreting social function and burial practice in the later Middle Ages. *Church Archaeology* **23**, 15 pages. https://doi.org/10.3828/churcharch.2023.23

[7] The Human Remains Project. 2024a. Retrieved 19 May, 2025 from Github HRDL: online_archives_consulted [https://github.com/HumanRemainsDL/HRDL/blob/main/online_archives_consulted.md]

[8] Paul Courtney. 2007. Historians and Archaeologists: An English Perspective. *Historical Archaeology* **41**, 11 pages. https://doi.org/10.1007/BF03377003

[9] David Gaimster and Roberta Gilchrist (Eds.). 2003. *The Archaeology of Reformation,1480–1580*. Routledge, London.

[10] Greg Walker. 2010. When did 'the Medieval' End? Retrospection, Foresight, and the End(s) of the English Middle Ages. In *The Oxford Handbook of Medieval Literature in English,* Greg Walker and Elaine Treharne (Eds.), Oxford University Press, 725–738. https://doi.org/10.1093/oxfordhb/9780199229123.013.0037.

[11] PAS. 2024. The Portable Antiquities Scheme: T. B. Periods used on the database. Retrieved from https://finds.org.uk/datalabs/terminology/periods/

[12] Richard Sharpe. 1997. *A Handlist of the Latin Writers of Great Britain and Ireland Before 1540*. Brepols, Turnhout.

[13] Michael Lapidge and Rosalind C. Love. 2001. The Latin hagiography of England and Wales (600–1550). in *Histoire internationale de la littérature part 3* 203–326. Brepols, Turnhout.

[14] David Farmer. 2011. The Oxford Dictionary of Saints. In *The Oxford Dictionary of Saints*. Oxford University Press, Oxford.

[15] Patrick J. Geary. 1991. *Furta Sacra: Thefts of Relics in the Central Middle Ages: Revised Edition*. Princeton University Press, Princeton, NJ.

[16] Goscelin. 2004. *Goscelin of Saint-Bertin: The Hagiography of the Female Saints of Ely*. Oxford University Press, Oxford.

[17] Alan T. Thacker and Richard Sharpe (Eds.). 2002. *Local Saints and Local Churches in the Early Medieval West*. Oxford University Press, Oxford.

[18] William of Malmesbury: Gesta Pontificum Anglorum. 2007. The History of the English Bishops Volume I: Text and Translation: 1. Oxford University Press, Oxford.

[19] The Human Remains Project. 2024b. HRDL: searchterms_all_languages Retrieved 19 May, 2025 from Github. https://github.com/HumanRemainsDL/HRDL/blob/main/searchterms_all_languages.md

[20] OED. exhumation.

[21] Jane Roberts, Christian Kay, and Lynne Grundy. 2017. A Thesaurus of Old English. Retrieved from https://oldenglishthesaurus.arts.gla.ac.uk/.

[22] Andy Orchard. 2022. *Word-Hoard: A Lexicon of Old English Poetry*. CLASP, Oxford.

[23] The Human Remains Project. 2024c. HRDL: variantrules_welsh. Retrieved 19 May, 2025 from Github. https://github.com/HumanRemainsDL/HRDL/blob/main/variantrules_welsh.md

[24] Isabelle Gribomont. 2023. OCR with Google Vision API and Tesseract. *Programming Historian*. Retrieved 19 May, 2025 from https://programminghistorian.org/en/lessons/ocr-with-google-vision-and-tesseract/.

[25] DOE. 2024. The Dictionary of Old English (DOE). Retrieved 19 May, 2025 from https://doe.artsci.utoronto.ca/.

[26] Regents of the University of Michigan. 2025. Corpus of Middle English Prose and Verse. Retrieved 19 May, 2025 from https://quod.lib.umich.edu/c/cme/.

[27] Forum on Information Standards in Heritage. 2014 (updated 2024). FISH Terminologies. *Forum on Information Standards in Heritage.* Retrieved from https://heritage-standards.org.uk/fish-vocabularies/

[28] The Human Remains Project. 2024d. HRDL: metadata_capture_values. Retrieved 19 May, 2025 from Github. https://github.com/HumanRemainsDL/HRDL/blob/main/metadata_capture_values.md

[29] Andreas van Cranenburgh and Gertjan van Noord. 2022. OpenBoek: A Corpus of Literary Coreference and Entities with an Exploration of Historical Spelling Normalization. *Computational Linguistics in the Netherlands Journal* **12**, 16 pages.

[30] Marc Kupietz, *et al.* 2020. *Recent Developments in the European Reference Corpus EuReCo*. Presses universitaires de Louvain, Louvain.

[31] The Human Remains Project. 2024e. HRDL: Python Scripts : ref_to_anl. Retrieved 19 May, 2025 from Github. https://github.com/HumanRemainsDL/HRDL/blob/main/Python%20Scripts/ref_to_anl.py

[32] Dawn Archer, Merja Kytö, Alistair Baron, and Paul Rayson. 2015. Guidelines for normalising Early Modern English corpora: Decisions and justifications. *ICAME Journal* **39**, 19 pages. https://doi.org/10.1515/icame-2015-0001

[33] Anthony Sinclair. 2022. Archaeological Research 2014 to 2021: an examination of its intellectual base, collaborative networks and conceptual language using science maps, Internet Archaeology **59**, 45 pages. https://doi.org/10.11141/ia.59.10

[34] Franco Niccolucci. 2017. Documenting archaeological science with CIDOC CRM. *Int J Digit Libr* **18**, 8 pages. https://doi.org/10.1007/s00799-016-0199-x

[35] Chiara E. Catalano, Valentina Vassallo, Sorin Hermon, and Michela Spagnuolo. 2020. Representing quantitative documentation of 3D cultural heritage artefacts with CIDOC CRMdig. *Int J Digit Libr* **21**, 14 pages. https://doi.org/10.1007/s00799-020-00287-3

[36] Cezary Mazurek, Krzysztof Sielski, Justyna Walkowska, and Marcin Werla. 2012. Applicability of CIDOC CRM in Digital Libraries. *CIDOC CRM Resources*. Retrieved from https://cidoc-crm.org/Resources/applicability-of-cidoc-crm-in-digital-libraries/.

[37] Nina Richards, Stefan Eichert, and Alexander Watzinger. 2022. One Ontology to Rule Them All—CIDOC CRM in the Humanities and Its Use in OpenAtlas. *goINDIGO,* 10 pages. https://doi.org/10.48619/indigo.v0i0.711.

[38] Martin Doerr. 2000. Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM. *ICS/FORTH, Technical Report* 274.

[39] Laís Carrasco and Silvana A. Vidotti. 2015. Dublin Core and CIDOC CRM Harmonization. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 3 pages. https://doi.org/10.23106/DCMI.952137181.

[40] Thanados. 2022. THANADOS: The Anthropological and Archaeological Database of Sepultures. Retrieved from https://thanados.net/.

[41] Roland Filzwieser and Stefan Eichert. 2020. Towards an Online Database for Archaeological Landscapes. Using the Web Based, Open Source Software OpenAtlas for the Acquisition, Analysis and Dissemination of Archaeological and Historical Data on a Landscape Basis. *Heritage* **3**, 16 pages. https://doi.org/10.3390/heritage3040077

[42] University of Cambridge. 2024. Cambridge Digital Library. Retrieved from https://cudl.lib.cam.ac.uk/.

[43] Arianna Ciula. 2020. Exposing legacy project datasets in Digital Humanities. *King's Digital Lab.* Retrieved 19 May, 2025 from https://kdl.kcl.ac.uk/blog/legacy-project-datasets/.

[44] Zeljko Carevic, Sascha Schüller, Philipp Mayr, and Norbert Fuhr. 2018. Contextualised Browsing in a Digital Library's Living Lab. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries,* 9 pages. https://doi.org/10.1145/3197026.3197054.

Just Accepted

[45] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Bee. 2018. A Study of Position Bias in Digital Library Recommender Systems. Preprint retrieved from arXiv.1802.06565.

[46] Matthew Reidsma. 2019. *Masked by Trust: Bias in Library Discovery*. Litwin Books, Sacramento, CA.

[47] Risto Ollila, Niko Mäkitalo, and Tommi Mikkonen. 2022. Modern Web Frameworks: A Comparison of Rendering Performance. *Journal of Web Engineering* **21**, 24 pages. https://doi.org/10.13052/jwe1540-9589.21311.

[48] Laura Morreale, *et al.* 2023. Digital Scholarship in the Field of Church History. *The Catholic Historical Review* **109**, 18 pages. https://doi.org/10.1353/cat.2023.0005.

[49] Xiying Mi, Richard Bernardy, and LeEtta Schmidt. 2021. Building an archaeological data repository: a digital library and digital humanities collaboration at the University of South Florida. *Int J Digit Libr* **22**, 10 pages. https://doi.org/10.1007/s00799-020-00297-1.

[50] Roger Louis Martinez-Davila, Sean Perrone, Francisco Garcia Serrano-Nebras, and Maria Martin de Vidales Garcia. 2020. Deciphering Secrets of Medieval Cathedrals: Crowdsourced Manuscript Transcriptions and Modern Digital Editions. *Bulletin for Spanish and Portuguese Historical Studies* **43**, 34 pages. https://doi.org/10.26431/0739-182X.1289.

[51] Arthur Burns, Kenneth Fincham, and Stephen Taylor. 2004. Reconstructing Clerical Careers: The Experience of the Clergy of the Church of England Database. *The Journal of Ecclesiastical History* **55**, 9 pages. https://doi.org/10.1017/S0022046904001514.

[52] Maristella Agosti, Nicola Orio, and Chiara Ponchia. 2018. Promoting user engagement with digital cultural heritage collections. *Int J Digit Libr* **19**, 353–366. https://doi.org/10.1007/s00799-018-0245-y.

[53] Annamaria Goy, Davide Colla, Diego Magro, Cristina Accornero, Fabrizio Loreto, and Daniele Paolo Radicioni. 2020. Building Semantic Metadata for Historical Archives through an Ontology-driven User Interface. *ACM J. Comput. Cult. Herit.* 13, 3, Article 25 (August 2020), 36 pages. https://doi.org/10.1145/3402440.

[54] Chern Li Liew. 2014. Towards dynamic and evolving digital libraries. *The Electronic Library* **32**, 14 pages. https://doi.org/10.1108/EL-07-2012-0089.

[55] Gustavo Candela, Javier Pereda, Dolores Sáez, Pilar Escobar, Alexander Sánchez, Andrés Villa Torres, Albert A. Palacios, Kelly McDonough, and Patricia Murrieta-Flores. 2023. An Ontological Approach for Unlocking the Colonial Archive. *ACM J. Comput. Cult. Herit.* 16, 4, Article 74 (November 2023), 18 pages. https://doi.org/10.1145/3594727.

Received 28 May 2025; revised 5 February 2026; accepted 18 April 2026

Just Accepted