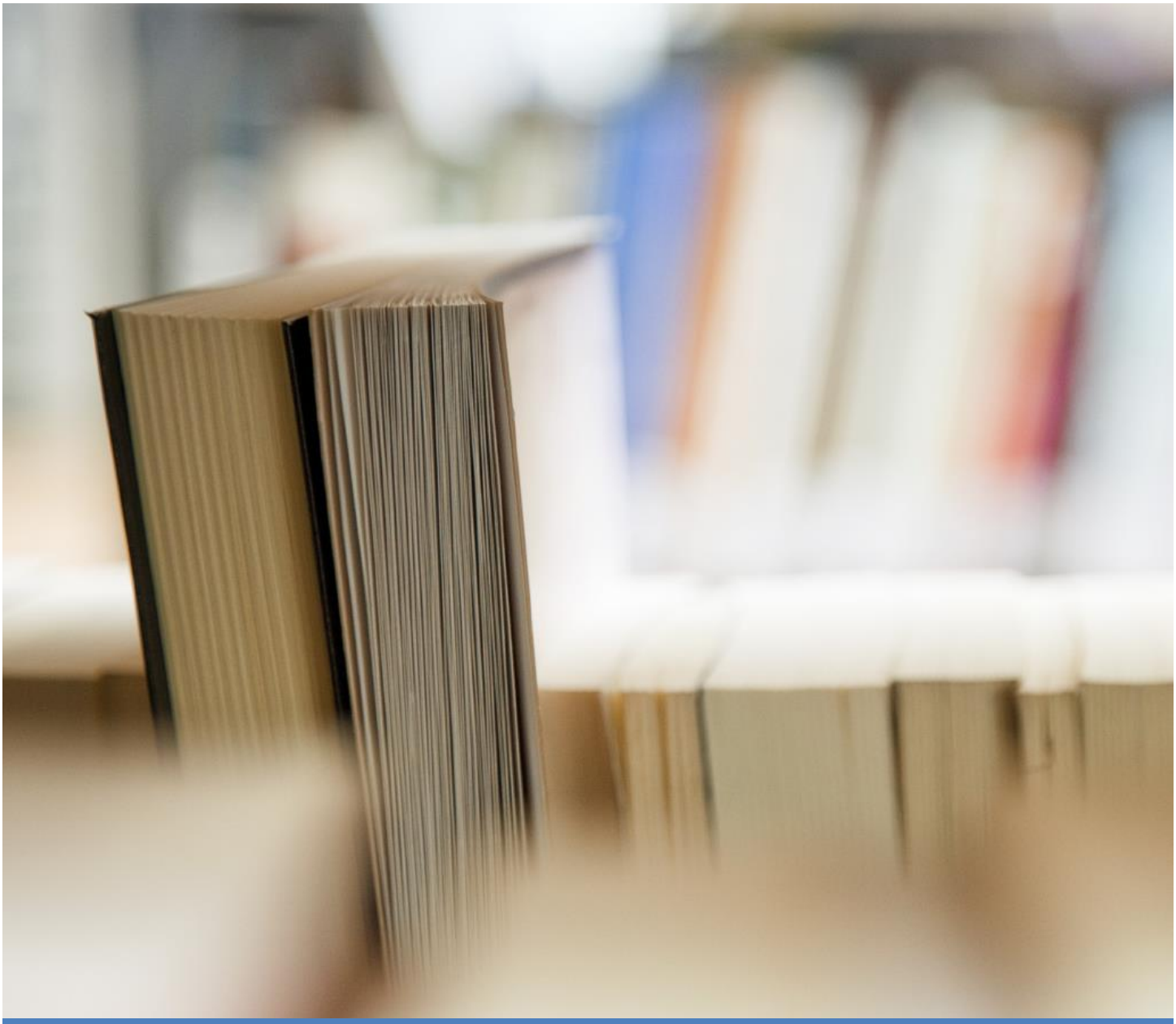


Towards a Global Dataset of Digitised Texts

Final Report of the Global Digitised Dataset Network

Towards a Global Dataset of Digitised Texts: Final Report of the Global Digitised Dataset Network

Lead authors (contact): Paul Gooding (paul.gooding@glasgow.ac.uk) and Natalie Fulkerson. Glasgow and Ann Arbor (MI) 2020.

Contributors: Paul Gooding (Principal Investigator) and Lucy Janes (University of Glasgow); Mike Furlough (Co-Investigator), Heather Christenson, Natalie Fulkerson, Joshua Steverman and Martin Warin (HathiTrust); Torsten Reimer, Alan Danskin, Tanya Kirk and Amelie Roper (British Library); Stuart Lewis, Sarah Ames, Ines Byrne and Gill Hamilton (National Library of Scotland); Dafydd Tudur, Siôn England and Owain Roberts (National Library of Wales); Matt Greenhall (RLUK).

About this document

This report describes the findings of the *Global Digitised Dataset Network* project (2019-2020), which investigated the feasibility of developing a single global dataset documenting the extent of digitised works. It sets out two key areas of work - identification of core use cases, metadata aggregation and data matching – and identifies a clear value proposition for the development of the proposed dataset.

We are grateful to all workshop attendees for their support and input, and for the enthusiastic response that we received from colleagues across the library sector.

The research that informs this white paper was supported by a grant from the Arts and Humanities Research Council under grant number AH/S012397/1 between February 2019 and January 2020.

Contents

Introduction

This report is one output of the AHRC-funded Global Digitised Dataset Network (GDDNetwork)[1]. The network was set up in response to a research networking highlight notice for UK-US collaborations focused upon digital scholarship in cultural institutions.[2] We set out to develop a new collaboration, led by the University of Glasgow and HathiTrust, to investigate the feasibility of developing a single global dataset documenting the extent of digitized works. The core network partners also include the British Library (BL), the National Library of Wales (NLW), the National Library of Scotland (NLS), and Research Libraries UK (RLUK). The network ran from February 2019 to January 2020, and aimed to:

- Seek to answer the question of whether it is feasible and worthwhile to create a global dataset of digitised texts for digital scholars, libraries, and readers;
- Develop a stronger understanding of the potential impact of a global dataset of digitised texts, particularly in relation to supporting digital scholarship;
- Investigate models for developing a sustainable global dataset, expanding on the network's initial scope to develop a truly global network.

This report will explore the work of the network, and establish priorities for refining and developing a sustainable model. We will provide an overview of existing aggregated online discovery platforms, in order to establish how a new platform might differ, and the benefits that those differences would bring to librarians and scholars. We will then expand upon the work of the GDDNetwork, demonstrating the results of two key work packages: the identification of key use cases from data gathered via workshops and surveys; and the development of automated metadata matching techniques to support collections overlap analysis. The report will conclude by proposing future priorities for continuing the development of a global dataset of digitised texts.

The project has offered the opportunity for us to reflect on the unique value proposition that a global dataset of digitised texts would offer. We have identified three key areas where the proposed dataset would present a unique and valuable contribution to the library ecosystem:

1. It would provide a platform for web-scale discovery of digitised texts, and support computational research by facilitating the identification of suitable digitised texts from around the world.
2. It would support collective strategic decision-making around digitised collections, and allow organisations to make local collection development decisions based upon comprehensive global collections data.
3. It would provide a focal point for a global network of practitioners and researchers to address issues of representation and diversity in trans-national collections.

While some aspects of this proposition are present in other resources, we will demonstrate that no single resource currently achieves these objectives at a global, and comprehensive, scale. We hope that the report will act as a catalyst for further research and development towards realising the value of the proposed dataset, in a manner that would support collective approaches to maximising the benefits of digitised materials.

1 Further details on the scope and work of the project can also be found online at https://gddnetwork.arts.gla.ac.uk/.

2 The specific details of the call are available at: https://ahrc.ukri.org/funding/apply-for-funding/archived-opportunities/research-networking-highlight-notice-for-uk-us-collaborations-in-digital-scholarship-in-cultural-institutions/.

Research Context

Libraries have been engaged in the mass digitisation of their collections for over twenty years, through local efforts, collaborations with other libraries, and partnerships with third parties such as Google, Microsoft, Gale, the Internet Archive, and Adam Matthew Digital. The project partners of the GDDNetwork are similarly active. HathiTrust, for instance, provides an aggregate discovery point for the mass-digitised collections of many libraries, and holds the largest collection of Google-digitised volumes outside Google itself. The British Library, National Library of Scotland, and National Library of Wales each come from the perspective of hosting large-scale digitisation programmes, and hosting and sustaining existing digitised collections. RLUK, meanwhile, acts to shape the research library agenda, and in recent years has published reports on digital scholarship, digitisation, and collaboration between sectors. This work sits within a broader recognition of the potential value of digitised library collections.

Having taken inspiration from the immense promise of digital collections for research, and the opportunity to derive new public and economic benefits from their collections (Niggemann, De Decker, & Levy, 2011, p. 4), libraries are now increasingly turning their attention to initiatives that facilitate scholarship through access to digital materials and, increasingly, services that support emerging forms of digital scholarship. Despite significant steps forward in supporting access and reuse, it remains extremely challenging to locate digital sources, particularly those produced by mass digitisation programmes. For example, Google has scanned tens of millions of books from over 20 libraries in North America, Europe, and Asia, but there is no publicly available index listing the titles that are available from Google Books or individual libraries. While users may employ a full-text search online, only Google knows the extent of what Google has scanned, not the libraries or scholars that nevertheless benefit from this work. The same opaqueness applies to the increasing amount of items created through libraries' own mass digitization programmes. A registry of digitised texts would promise to improve discoverability, and to allow libraries and scholars alike to view library holdings at web scale. There have been many efforts to improve discoverability of digital resources in general, and digitised resources specifically. These efforts can be divided into three categories based on their scope:

1. **Local or project-based**: focused upon a single digitised collection (e.g. The Burney Collection of Newspapers at the British Library) or set of digitised collections (e.g. Welsh Newspapers Online by the National Library of Wales) that are hosted by a single organisation or content provider.
2. **Intra-national:** focused upon a single set of digitised collections within a single nation or legal jurisdiction. This includes format or domain-specific platforms (e.g. the National Bibliographic Knowledgebase), and national-scale multi-format platforms (e.g. Digital Public Library of America).
3. **Trans-national:** focused upon a set of digitised collections from multiple nations or legal jurisdictions. This includes international format or domain-specific platforms, and large-scale multi-format aggregation platforms (e.g. Europeana Collections).

However, significant problems remain for scholars undertaking digital scholarship, and for libraries engaging in supporting digital scholarship as broadly defined. The recent RLUK report into digital scholarship in research libraries draws upon Melanie Schlosser's definition of digital scholarship as "research and teaching that is made possible by digital technologies, or takes advantage of them to ask and answer questions in new ways" (2012). Her definition is deliberately vague, and can therefore encompass several aspects of research and teaching facilitated by digital technologies: the application of computational methods to research the exploitation of digital technologies in the research cycle, and the multi-disciplinary nature of digital scholarship (Greenhall, 2019, p. 10). Such a broad definition necessitates consideration of the utility of discovery platforms across the whole content lifecycle.

Several reports over the past fifteen years have noted the potential benefits of developing trans-national discovery platforms for scholarship and culture. In 2005, for instance, the Commission of the European Communities noted that social and cultural benefits could accrue by representing the richness of the world's history, and its "cultural and linguistic diversity" (2005). Indeed, Nanna Bylde Thylstrup argues that mass digitisation has played a key role in supporting the development of a form of "national imaginary" (2019, p. 58) by allowing countries to promote online a sense of shared national identity. In this way, national cultural and political priorities are fulfilled by digitisation. For instance, the National Library of Scotland describes "small, smart nations," an informal community of interest amongst countries that are technologically, politically, educationally, and socially advanced but are too small to be global leaders (Hunter and Brown, 2010). Such countries have derived benefits from collaborative efforts to promote and make available their cultural heritage in order to protect distinct national identities from globalisation.

Similarly, the economic benefits of digitisation as a product of the cultural heritage industry should not be underestimated. In a report for Historic Scotland, it was noted that the English cultural heritage sector contributed to a Gross Value Added to the UK's Gross Domestic Product of £13.1 billion in 2016 (Cebr, 2018). Digitisation is seen as a vital component in realising the value of cultural heritage that turns heritage materials into "a formidable asset for the individual user and an important building block of the digital economy" (Niggemann, De Decker and Levy, 2011, p. 4). Similarly, the UK government's Digital Strategy 2017 (Department for Digital, Culture, Media & Sport, 2017) and Culture is Digital Report (Department for Digital, Culture Media & Sport, 2018) both emphasise the importance of digitisation in realising continued benefits from heritage organisations. The Europeana Collections initiative ties into Europe-wide efforts to harmonise copyright and increase interoperability between Information and Communications Technology systems in support of the digital single market (European Union, 2017). The resultant benefits are broad: in academic research, teaching, and lifelong learning; in leisure and tourism; in commercial exploitation of open cultural content; and in the diverse network of jobs that support the delivery of cultural heritage materials online. These diverse indicators of social, cultural, and economic good provide an important justification to increase the discoverability and reuse of digitised texts.

Bilansky (2016, p. 2) points out that the shift to online discovery of cultural heritage has transformed citation patterns, reading habits, and research methods. However, locating these sources, and particularly large corpora of digitised texts, remains difficult: rights issues continue to restrict access after digitisation; statutory and contractual restraints limit sharing; and despite significant digitisation activity, many countries continue to report a lack of overarching national strategy (European Commission, 2014, pp. 7–9). Furthermore, organisational policies have, in many cases, not caught up with the emerging demands of digital scholarship, and the interfaces that support discovery and reuse often focus upon viewing items one at a time in a browser, rather than allowing users to embrace the myriad possibilities of digital artefacts (Whitelaw, 2015). Commercial providers such as Gale have taken steps to develop interfaces that incorporate tools for digital scholarship,[3] but licensing and access restrictions in both commercial and non-profit organisations mean that these benefits are limited to specific collections and platforms. As a result, many digitisation efforts do not achieve the impact upon digital scholarship that they might do with a more coherent digitisation policy.

Even those platforms that provide vital services to the library community only partially fulfil our vision of a resource that supports discovery, computational analysis with, and of, library holdings, and systemic decision-making in the library sector. A registry of digitised texts would be tens of millions of records large, and to the extent that it represents books or other textual formats that are conventionally described using MARC, one may wonder whether OCLC's WorldCat service might not substitute for, or support, this work. As of December 2019, WorldCat includes 470 million records for 2.8 billion book or serial holdings worldwide (WorldCat, 2020). Despite appearances to the contrary, this is neither comprehensive nor truly representative of worldwide

[3] See, for instance, the Gale Digital Scholar Lab, which offers common Digital Humanities tools to researchers using Gale collections: https://www.gale.com/intl/primary-sources/digital-scholar-lab.

collections. Holdings of libraries in the UK, for example, have been historically under-represented in WorldCat, although this has recently begun to change. In many instances, WorldCat records the existence of digital versions of a text, either as a separate edition of the printed work, or as a manifestation of that printed work. For example, HathiTrust is typically listed as a "holdings library" for nearly 9 million titles that have been digitised and deposited in HathiTrust as of December 2019, for which users can obtain access. However, especially when the source library for the HathiTrust record is only partially included in WorldCat, or not at all, those digital copies cannot be located using the catalogue. In another example, the British Library records for included in our pilot project had been exported to WorldCat, but for reasons that are not clear, WorldCat does not always indicate that these items have digital manifestations available for online access. These inconsistencies limit access and discovery by scholars and information professionals alike, and indicate the need for focused attention to documentation and discovery of digitized materials. Finally, while OCLC has service offerings to help libraries de-accession physical volumes, and uses the existence of digital version as a data point in such work, it does not make the WorldCat data available in a format that is actionable to support library digitisation choices and strategies. An alternative solution is therefore required to ensure comprehensiveness, and to address discovery, use, and computational reuse, of digitised library collections.

In order to solve these challenges, it has been pointed out that the library sector must begin to rethink its assumptions around technical infrastructures and cultures of collaboration. In 2013, for instance, Dempsey et al. noted "libraries and the organizations that provide services to them will devote more attention to system-wide organization of collections – whether the "system" is a consortium, a region or a country" (Dempsey *et al.*, 2013, p. 1). Shared approaches to large-scale discovery of collections have been evident in national and trans-national collaborations such as the Digital Public Library of America and Europeana Collections, and there is an overall shift towards understanding that access and discovery of digitised collections are systemic problems rather than local ones. Furthermore, there is a broad trend towards sharing collections via an increasingly diverse range of platforms; for instance, many organisations have shared their digitised materials under open licenses via Wikimedia Commons and Europeana Collections. These initiatives often focus upon improving conditions for discovery and reuse of digital collections for users. However, collective solutions to these challenges also promise to transform strategic policy development in the library sector, through improved sharing of collections and related data between organisations. Dempsey uses the term "collective collection" to describe the "collective development, management and disclosure of collections across groups of libraries" (Dempsey, 2013, p. 1) that are demonstrated in these activities.

Existing Large-Scale Discovery Platforms

Several initiatives exist to support large-scale discovery of cultural heritage materials. This section focuses on three further platforms, in order to identify gaps in existing provision, and explore how a possible dataset that documents the global extent of digitised works might provide distinct but complementary services. It will look at the Digital Public Library of America, the UK National Bibliographic Knowledgebase, and Europeana Collections; each representing a national or trans-national approach to the challenge of developing strategically aligned collective collections. For each, we will outline the history and development of the platform, and refer to publicly available sources to establish key strategic aims and objectives.

Digital Public Library of America

Outline

Members of the Berkman Center founded the intra-national Digital Public Library of America (DPLA) in 2010 for Internet and Society at Harvard University. It initially received funding from the Alfred P. Sloan Foundation, and secured subsequent funding from other sources including the National Endowment for the Humanities and the Bill & Melinda Gates Foundation. The DPLA described its founding vision as follows:

> A single point of access. A gateway to the cultural and intellectual riches stewarded by libraries, archives, and museums across the United States. Open to all, and only possible in the digital era (DPLA, 2019).

This vision was applied to the creation of an open, distributed national library with the aim of informing and empowering everyone. Beginning in 2011, a two-year process was launched to bring together stakeholders from libraries, academia, innovation, and the volunteer community, led by a steering committee to develop six work streams to scope, design, and construct the DPLA.[4] The DPLA then became an independent organisation in 2013 under the directorship of Daniel J. Cohen, succeeded in 2017 by John Bracken. As of 2019, the DPLA represented 4,000 contributing institutions from 41 US states, providing a discovery platform for over 30 million items.

Key Aims and Objectives

The DPLA states that its overarching mission is to make digital content from US libraries and cultural organisations accessible to all. To support this, the DPLA has two major strands running through its strategic roadmap. First, it foregrounds the philanthropic nature of the endeavour, through an explicitly mission-driven approach that aims to ensure that the stories of all Americans are told, and that all people are able to access those stories through the DPLA infrastructure. This combines a desire to record and promote US culture worldwide, with a complementary desire to ensure access and reuse of heritage materials from the entire nation. The focus of the collections are upon archival materials, although it does include over 3 million records of books found in HathiTrust. For cultural heritage materials, DPLA aggregates metadata only, while content remains hosted on partner and hub services. With this in mind, the strategic plan draws attention to three priority areas of work. First, it aims to make millions of materials available in a single discovery platform encompassing galleries, libraries, archives, museums, and cultural heritage organisations. Second, it aims to provide a library-controlled marketplace for purchasing and delivering eBooks and other digital content to patrons. Third, it aims to provide opportunities for library leaders and practitioners to “explore and advance technologies that serve, inform, and empower their communities” (DPLA, 2019, p. 2).

[4] The work undertaken during this phase is freely available via the DPLA wiki at https://digitalpubliclibraryofamerica.atlassian.net/wiki/spaces/HM/pages.

The DPLA strategic roadmap hints at services provided to libraries, but other than the marketplace, these elements have not been developed to the same extent as other services. Instead, the DPLA roadmap refers to a planned "full-service, library-controlled pathway to purchase, organize, and deliver eBooks and audiobooks" (DPLA, 2019, p. 1). It opens up the possibility of developing services for collaborative decision making in the future; in the short term, the intention appears to be to support local evidence-based decision making for prospective eBook acquisition and licensing. At this stage, the roadmap focuses primarily on providing access, and upon developing a national e-content service to partner organisations.

The UK National Bibliographic Knowledgebase

Outline

The UK National Bibliographic Knowledgebase (NBK) is the result of a project led by Jisc, to build a new UK-wide service to improve access to print and digital monograph resources and to help libraries collaboratively manage their collections. No single public document expresses the strategy behind the NBK, but it was originally proposed as part of the Jisc National Monograph Strategy Roadmap in 2014. This roadmap provides a broader strategic plan for monographs in the UK, and "describes the components of a collaborative, national infrastructure that provides answers to the past, present and future challenges of the scholarly monograph" (Showers, 2014). The NBK is a foundational requirement for the broader infrastructure, as it underpins several components of the strategic plan. In response to the roadmap, Jisc began to work on the development of the NBK, and OCLC was subsequently contracted to deliver the platform. The delivery phase of the project occurred between January 2017 and July 2019, and on 31st July the NBK service was taken out of beta. It offers three complementary "library hub services": "discover", a national scale discovery interface for UK academic library collections including monographs and journals;[5] "compare", which allows libraries to compare their collections to other contributors to the NBK;[6] and cataloguing, which provides a platform for producing and submitting catalogue data for contributors[7] (Jisc, 2019).

Key Aims and Objectives

The NBK is one component of a unified vision for a UK monograph strategy that comprises the following components:

- An "open, comprehensive, accurate and timely bibliographic and holdings knowledgebase" (Showers, 2014) that forms a core piece of monograph infrastructure for print and digital monographs;
- A service to enable researchers, libraries, and publishers to track and manage the impact of their monographs;
- A shared open monograph publishing platform that would provide a sustainable solution to lower the threshold for monograph publishing.
- The use of these services to inform the development and testing of new monograph business models, and a national strategy informed by evidence from the monograph knowledgebase; and a license negotiated by a third party to secure access to digital scholarly monographs for the UK academic sector and to reduce overhead costs for libraries and publishers.

The NBK aims to improve accessibility and discoverability of monographs, and to support innovative reuse for new forms of research, but its intended remit also extends to supporting and

5 See https://discover.libraryhub.jisc.ac.uk/.

6 This website is login only, but further information and support is available at https://compare.libraryhub.jisc.ac.uk/.

7 Again, this is a password-protected service, but the home page is available at: https://cataloguing.libraryhub.jisc.ac.uk.

establishing community-based library practices. It aims to allow the development of a collaborative national monograph strategy for the United Kingdom, and to support libraries to adopt systemic decision-making practices for defining local monograph spending priorities. Its intended benefits include supporting publishers and presses to realise the value of their monograph backlist through digitisation, and by developing business models for digital monograph (Showers, 2014). One point of difference from a potential system-wide digitisation strategy is that de-duplication is a low priority: indeed Findlay (2019) notes that it is less important for NBK than ensuring the quality of resources. Throughout the development process, however, collective library decision-making has been one of the key delivery priorities for the NBK. This is a point of departure from the Europeana Collections approach, which the next session will show prioritises discovery and organisational impact evaluation.

Europeana Collections

Outline

Europeana Collections is the product of significant pan-European collaboration on the digitisation and online accessibility of cultural heritage materials since around 2000. Between 2000 and 2005, the European Commission co-funded research projects and stimulated collaboration between member states to present cultural heritage online (European Commission, 2008). The lineage of Europeana Collections can be traced back to the Gateway and Bridge to Europe's National Libraries (GABRIEL), a showcase for 43 national libraries. Supported by European Commission funding, GABRIEL grew into the European Library, a "search engine and open data hub for library collections" that was launched in 2005 (Kenny, 2017). 2005 also saw six heads of state call for more European investment in the European Library, supported by the national libraries of 19 European nations. José Manuel Barroso, the President of the European Commission, publicly welcomed the digital libraries initiative. The commission subsequently adopted the i2010: Digital Libraries strategy,[8] which outlined plans to develop and support a European Digital Library. The European Digital Library Foundation prototype went live in December 2008. It was subsequently rebranded Europeana, which became an operational service in 2010. In 2015, Europeana's collection-related websites were all brought together under a single service called Europeana Collections (Kenny, 2017). As of 2020, Europeana Collections provides access to over 58 million digital objects from a diverse community of European cultural organisations.

Key Aims and Objectives

This section bases its interpretation of the Europeana Foundation's key aims and objectives upon two documents: The Europeana Strategy 2015-2020 (Europeana, 2014), subsequently updated by the Europeana Strategy 2020 Update (Europeana, 2017). As noted above, Europeana Collections was envisaged as a way to unite European cultural heritage in a single portal available online for all. This was underpinned by a desire to realise the economic, social and cultural value of European culture. The strategic report and its update formalise and structure the work to achieve this aim, by clearly defining key stakeholder groups, the structure of Europeana's work, and the strategic priorities.

Europeana focused upon digitised content, with the aim of making it more usable for work, learning, and leisure. The Europeana Strategy notes that the initial aim of developing a single access point for library, museum, and archival content for Europe, while important, is not ambitious enough to allow for the multiple forms of reuse and interaction that audiences require (Europeana, 2014, p. 10). Since 2008, Europeana has developed around three areas of work: the development of a single point of access for all resources via www.europeana.eu; the application of Creative Commons Public Domain marks to metadata, in an attempt to share collated metadata as broadly as possible; and the development of the Europeana Data Model (EDM), which aims to

8 The full communication is available online at the following link: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52005DC0465

provide a single authoritative data model for participating institutions to submit consistent metadata. The EDM provides a framework for "collecting, connecting, and enriching metadata" (Europeana, 2015) and to be included in Europeana, data must be provided in either EDM, or the older Europeana Semantic Elements standard.[9] These three work strands have occurred alongside extensive user engagement with the objective to enable researcher reuse of heritage materials, and to ask users to share personal stories linked to particular historical events.

The next phase of the Europeana strategy is summarised by the phrase "From portal to platform" (Europeana, 2014), which aims to increase participation by enabling users to reuse materials and interact with other users. Europeana intends to define this platform in terms of a three level structure: a core level, where the platform collects data, content, and technology; an access level where Europeana standardises and enriches the core, and provides the interface and rules of access; and a service level where tailored services are created. These services will be targeted at heritage professionals, end users, and creatives. Finally, this shift towards a platform model has three priority areas. The first stated priority is to improve data quality: the aim here is to maintain a low entry threshold for contributing organisations, while developing infrastructure to surface higher quality materials and openly licensed content. Additionally, Europeana intends to shift towards a distributive data delivery architecture. This priority aims to triple the amount of content available through Europeana while foregrounding high quality materials that are suitable for innovative reuse (Europeana, 2014, p. 12). The second priority is to open the data so that it may be viewed and reused to the maximum extent allowed by copyright. A key pillar of this priority is to make all digitised Public Domain material freely available without restriction. The platform will continue to promote and use open metadata, and to provide a source of authentic, trusted cultural heritage materials (p.14). The third priority is to create value for partners. This is envisaged in terms of simplifying participation, providing better statistics to track visibility of content within Europeana, and developing opportunities for networking and development.

In summary, the objectives of Europeana are to foreground high quality digitised cultural heritage content; promote open content; support impact evaluation; and make open metadata available to aid discovery. The key point of difference from platforms such as NBK is that decision making is largely explained at a local level, rather than in terms of collective strategic development. There is therefore great scope to consider the potential benefits of a dataset that would address discovery and collective decision making. The following section addresses the work of the GDDNetwork to date. In light of the discussion above, this work addresses both the development of unique and valuable use cases for the dataset, and the related question of metadata matching for the purpose of de-duplication.

[9] The Europeana Sementic Elements Standard (ESE) was the original Europeana data model. It was replaced in 2013 by the EDM, which was considered a richer model that allows "explicit links between Europeana objects" (Europeana, 2013).

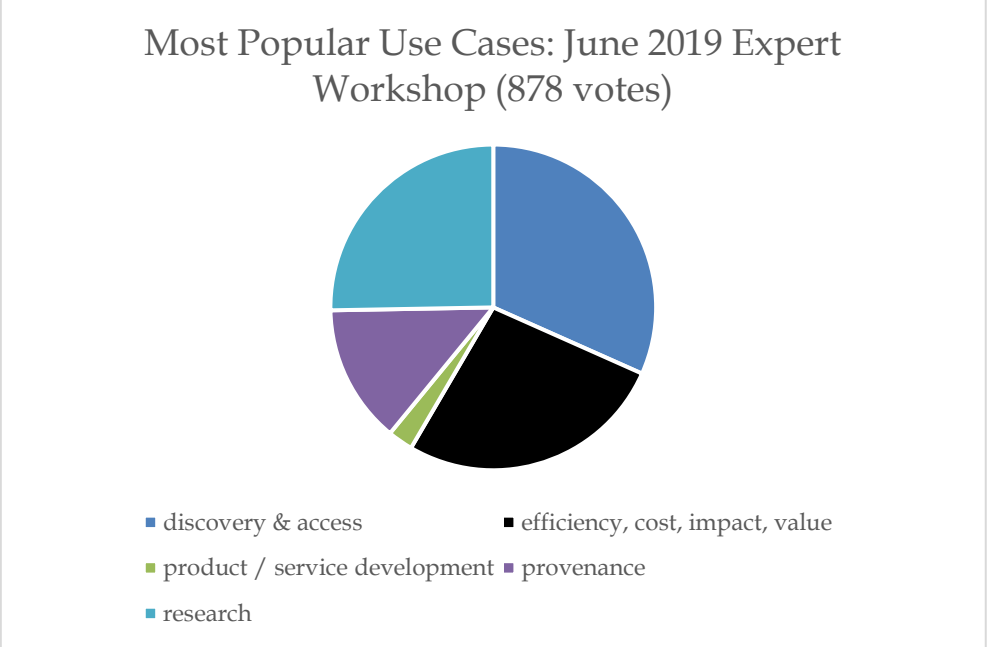
The Work of the GDDNetwork

Developing Use Cases for the Global Digitised Dataset

One of the key work strands of the project has been to identify and develop use cases to indicate the possible benefits and scope of a global dataset of digitised texts. As stated, the project set out with an initial idea of what a global dataset might offer, including supporting scholars in discovery and data-driven research, and allowing libraries to develop new strategic approaches to collaborative digitisation and efforts to understand collections overlap. However, we felt that it was important to establish the kinds of use cases that the stakeholder community could envisage for this resource, and to identify whether these use cases were feasible and sufficiently distinct from existing resources to make a standalone global dataset viable. It is worth noting that we were eager to explore and surface various possibilities, rather than to provide a rigorous process for identifying and prioritising use cases. The work we undertook was similarly exploratory in nature, and it should be borne in mind that further work is required to ensure the dependability of these findings beyond our small and unrepresentative samples. Taken in this spirit, our findings hint at future possibilities rather than a robust roadmap.

We approached the development of use cases in two ways. The first method was via brainstorming and workshopping ideas with key stakeholders. First, the core network team brainstormed agile user stories, adopting the format "As a *…* I want to *…* so that I can *…*". This gave us a longlist of possible use cases, which we categorised according to the following broad themes: discovery and access; efficiency, cost, impact, value; research; provenance; and product/service development. For each category, use cases were developed relating to a variety of key stakeholders: readers; digital scholars; collections managers; reader services managers; institutional leaders; metadata specialists; digitisation specialists; and vendors. In June 2019 we ran a subsequent workshop, inviting a group of expert stakeholders from relevant communities to brainstorm further use cases. Once we had established a broad longlist of use cases, we asked attendees to undertake an "investment" exercise: each participant was given 30 stickers, and invited to vote for their preferred use cases in order to establish a shortlist of priority investment areas.

In total 878 votes were cast. The following chart shows the proportion of votes for each category:

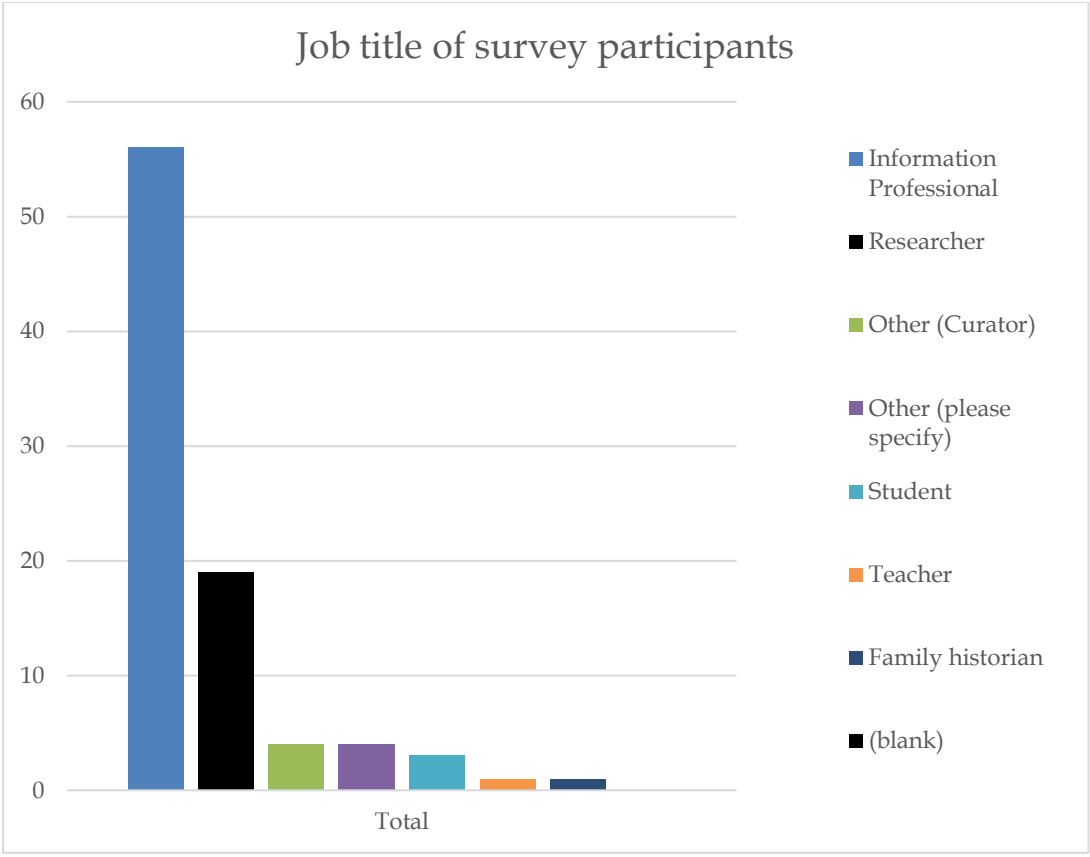


Attendees prioritised use cases relating to discovery and access, and efficiency, cost, impact and value. There was also a slight bias towards research-related topics (everything but the orange and grey areas above), and a significant lack of interest in product/service development cases. It should be noted, however, that this was partly because library service providers have been underrepresented in the network to date, in comparison to researchers and librarians supporting digital scholarship.

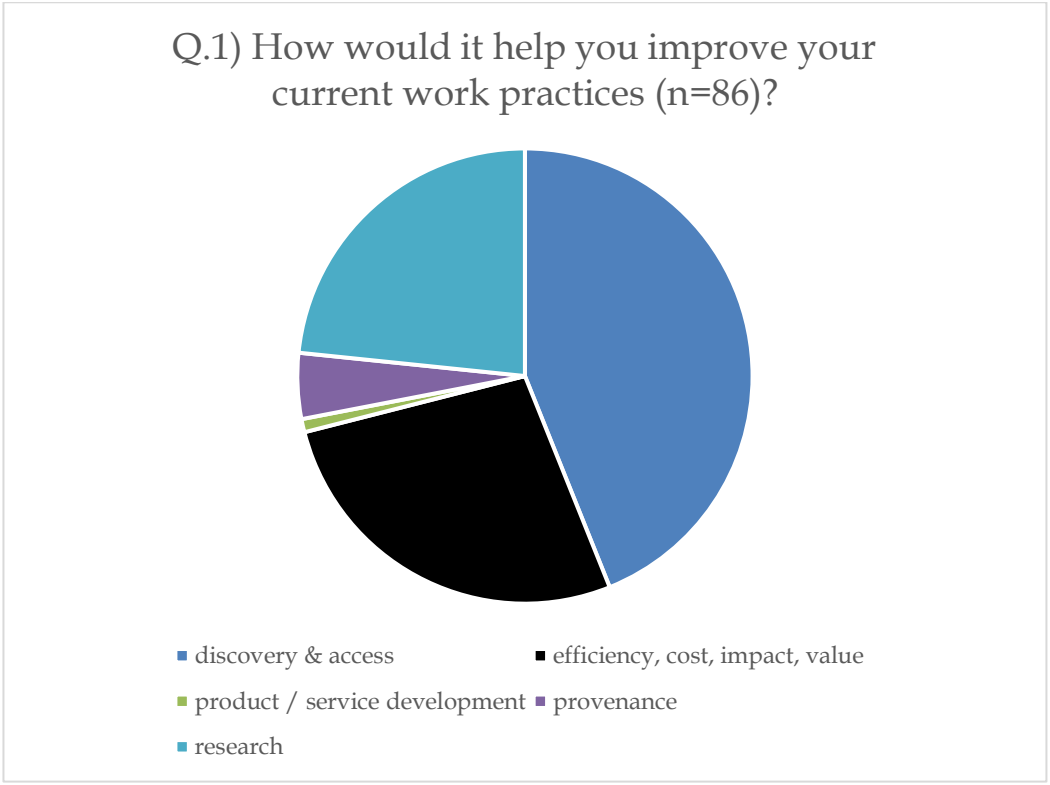
The following table demonstrates the most popular use case for each category:

Theme	Use case	Number of votes
Discovery and Access	"As a reader I want to easily, remotely access a digital resource so that I can find the information I'm after."	88
Efficiency, Cost, Impact, Value	"As a collections manager I want to know what has already been digitised so that I can avoid duplication of effort."	48
Provenance	"As a digital scholar I want to understand the provenance of the dataset so that I can put the digitised materials in context and apply my own relative score to the source (e.g. how much I trust it)."	31
Research	"As a digital scholar I want to download a list of links to digitised texts from different libraries so that I can create a corpus specific to my needs."	42
Product/Service Development	"As a vendor, I want to know what libraries have digitised so that I can include a new discovery channel in my product."	16

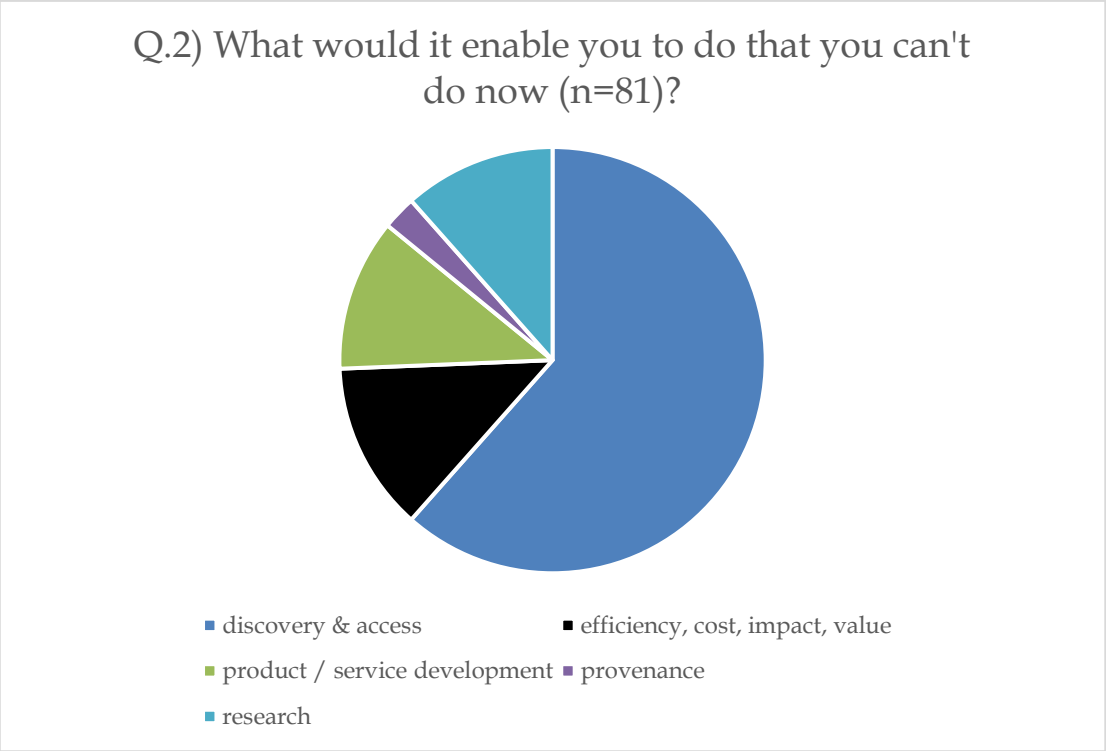
Our second approach was to undertake a survey of key stakeholders. The survey was developed collaboratively to gather evidence of the potential benefits of a global registry of digitised texts. As we wanted to establish which use cases would be most beneficial to the stakeholder communities, we explicitly focused upon questions that uncovered perceived benefits based upon a brief description of the intended resource. Some responses queried this, and so it is important to note that the results should be taken as indicative of the use cases that respondents thought would be beneficial, rather than as a proxy measure of support for the concept. The survey received ethical clearance from the University of Glasgow, and was delivered using the university's online survey platform. The survey was open from the 28th July 2019 to the 11th October 2019, and received 86 complete responses. The survey contained four questions, including one to establish the job category of respondents. The remaining three questions asked respondents to submit free text responses to questions on how the dataset would support working practices for them as individuals, and those in similar roles. These free text responses were coded by hand to reflect the same categories as those established in the workshops, and to identify additional categories that emerged. As the following table shows, the majority of respondents were information professionals or researchers:



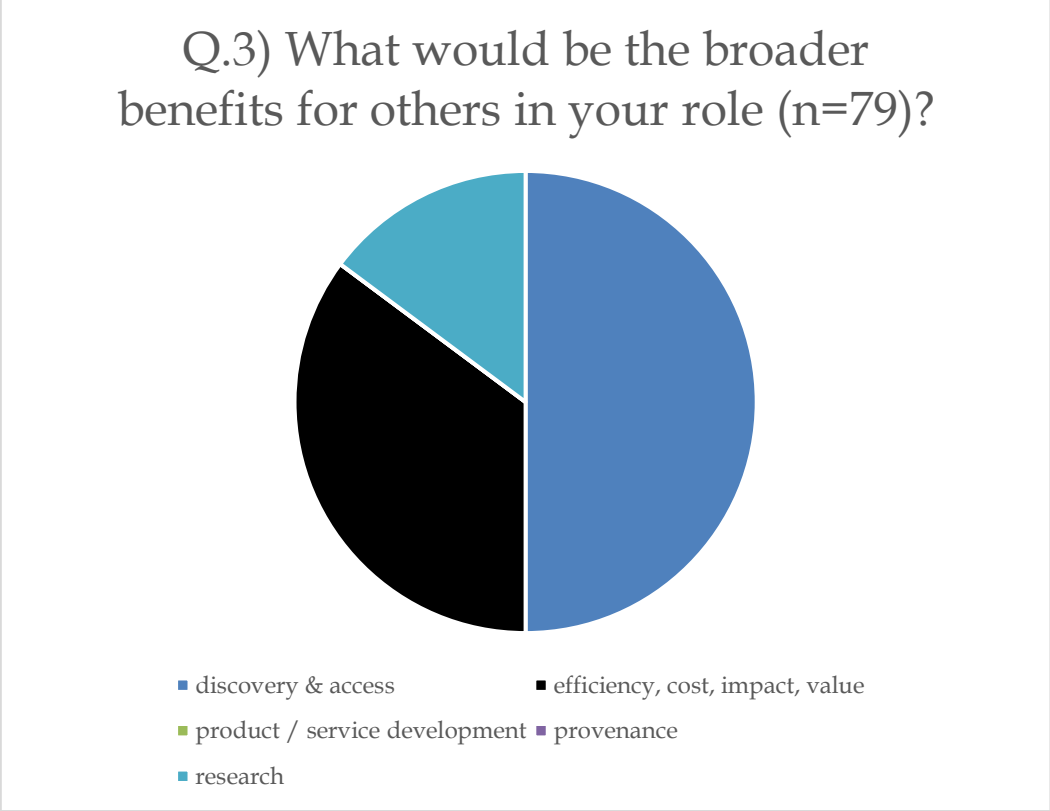
We found that, like our workshop attendees, survey respondents also prioritised "discovery and Access" and "efficiency, cost, impact, value". The following chart shows the most popular categories of use cases when respondents were asked about how it would improve current working practices:



When asked what the resource would allow them to do that they couldn't now, respondents focused more strongly upon "discovery and access":



Similarly, when asked to consider the impact upon those in similar roles, discovery and access became increasingly strongly represented:



In addition to the five key categories, several survey responses raised additional theoretical and critical themes. Several respondents focused on the idea of "democratisation of

knowledge" as a way of ensuring not only equality of access but increased consistency across research communities. One respondent commented, for instance that "IF there was access for all to, for example, the first edition of Dickens texts, and that could be used by everyone when referencing, that would be very helpful for consistency." Others noted issues of "digital preservation and sustainability". One respondent noted that a directory could "potentially de-duplicate efforts or help set priorities" for digital preservation efforts. Similarly, another respondent noted that any such registry would need to be distributed to ensure protection against large-scale data loss. A few respondents noticed key challenges that they believed should inform future work. Data formats and user interfaces were a particular concern. One respondent noted, for instance that "the biggest obstacle to overcome… is data formats and whether, having unified all digital texts through one portal, it would then be possible to compare them through an interface that handles different digitised text formats."

What therefore emerges from the case study work are three key benefits that might shape and define further thinking on the purpose of the registry: "discovery and access", "efficiency, cost, impact, value", and "research." "Provenance" was commonly referenced in relation to specific use cases, many of which were essential to support work in these areas but when asked to extrapolate generalised benefits the three headline categories consistently emerged. This suggests that use cases relating to "provenance" and "product/service development" might be viewed through these priority lenses. A key conclusion is therefore that further work should be done to define a mission statement based upon the three headline benefits, but encompassing key use cases relating to provenance and product/service development that would support these objectives.

A key finding of both approaches was that further work is required to define the scope and extent of the proposed dataset. Survey respondents in particular were slightly confused by the concept. Many respondents proposed case studies that were based on the assumption that it would be possible to provide direct access to digitised full text files. While this suggests an important next step, it is worth noting that the work of the network to date has been primarily on unifying metadata to support discovery and access to computationally reusable texts, not upon aggregating full text. This suggests, should access to full text fall into scope, that work on copyright and licensing issues will be key to the successful delivery of relevant use cases. Additionally, several respondents raised valid concerns about what it meant to create a "global" dataset: several noted that diversity in languages, cultural backgrounds and geographical reach were essential; others queried whether there would be an adequate balance between large and small organisations; and some expressed concerns about maintaining sufficient data quality to support effective holdings analysis.

Holdings Analysis

Our second priority was to develop a prototype registry of digitised texts using library catalogue data supplied by the four library partners (British Library, National Library of Scotland, National Library of Wales, and HathiTrust). This work was led by HathiTrust from February to July 2019. A significant component of this work was to design and test a methodology for identifying overlap between the four collections, in order to assess how well the use case pertaining to digitisation, collection management, research, and provenance could be met by such a registry. In particular, we set out to identify and evaluate various methods for determining overlap, gain experience working with unfamiliar records, and consider whether and how those methods could be applied at web scale. In a subsequent component, we aggregated catalogue data from all four library collections to produce the prototype registry.

OCLC Control Number (OCN) Matching

To begin with, each of the four libraries contributed full MARC records for their digitized holdings: the British Library contributed records for their 19th century monographs collection, and the NLS[10], NLW[11] and HathiTrust[12] provided digital holdings. The NLS and NLW also provided records for their print holdings.

	# of digitized records	**# of print records**
British Library	516,212	-
National Library of Scotland	10,919	9,640,360
National Library of Wales	2,290	3,224,243
HathiTrust	16, 987, 842	-

HathiTrust has an existing process for identifying overlap among its members' holdings that relies on the presence of an OCLC Control Number (OCN), which uniquely identifies each work, to determine whether more than one library holds a copy of the same title. OCLC's WorldCat aggregates holdings across participating libraries in a similar fashion. OCN assignment is more-or-less standard practice in US cataloguing, but UK libraries use them less often, so we suspected that most of the UK records would lack OCNs. As expected, very few of the UK records contained an OCN, rendering this first matching approach ineffective for identifying overlap.

10 The National Library of Scotland data was not a complete list of all digitised texts, which numbers significantly higher. It was based on a historical export of bibliographic data, and only contains records for monographs.

11 The National Library of Wales provided title-level records of all print works digitised by the Library.

12 The HathiTrust record set is, itself, an aggregation of catalogue data provided by the member libraries who contribute digital content to its collective collection.

	# records	# OCNs	# matching	% matching
British Library	516,212	611	130	0.025
National Library of Scotland	10,919	561	243	2.22
National Library of Wales	2,290	744	101	4.41

ISBN matching

Next, we searched the records for ISBNs, as a proxy for other identifiers.

	# digitized records	# print records	# ISBNs - digital	# ISBNs - print	% ISBNs - print
British Library	516,212	-	34	-	-
National Library of Scotland	10,919	9,640,360	55	2,709,837	28
National Library of Wales	2,290	3,224,243	17	3,128,171	97*
HathiTrust		-	2,645,141	-	16

We found that the presence of ISBNs was highly variable across both the UK and US records, eliminating ISBN as a reliable match point.

Exploratory approaches

Our next step was to define a methodology for applying more exploratory approaches to record matching. We adopted a 4-part framework: 1) identify methods, 2) pilot each method on a small subset of records, 3) evaluate each method, and 4) apply to the full set of records, if appropriate. We identified 3 exploratory methods, including title string matching and two statistical approaches; one using word-by-word (or “bag of words”) matching, and the other utilizing machine learning techniques.

String matching

We explored two title string matching approaches: raw matching and normalized matching. In each case, we extracted title fields from the 3 UK library records and compared them to the title fields in the HathiTrust records. The raw matching approach required an exact character-by-character match as expressed in the original record; in the normalized approach, we downcased the titles and removed non-alpha characters prior to matching them, to account for non-meaningful differences in casing, spacing, and punctuation. Results are shown in the following table.

Raw match:

	# digitized records	# matches	# matches to multiple records	% overlap
British Library	516,212	4,559	2,255	0.88
National Library of Scotland	10,919	343	131	3.14
National Library of Wales	2,290	51	9	2.23

Normalized match:

	# digitized records	# matches	# matches to multiple records	% overlap
British Library	516,212	39,815	18,298	7.71
National Library of Scotland	10,919	1,746	837	16
National Library of Wales	2,290	253	83	11.04

We saw more overlap using the normalized approach, which makes sense given that the normalizing process has the effect of making the records more similar to each other. For the same reason, we also saw an increase in the number of UK titles that matched to more than one HathiTrust title. More work is therefore needed to adequately investigate false negative and false positive rates.

Statistical approaches

In contrast to the binary ("match" or "no-match") approaches described above, word-by-word matching and machine learning methods cannot definitively identify overlap using individual record pairs. Rather, these methods are applied to the whole set of records at once, and generate results that are subject to statistical interpretation. Therefore, the findings described below do not attempt to represent overlap between entire collections. Instead, they are presented as measures of the relative success of each method.

Word-by-word matching

Word-by-word matching decouples character strings (in this case, words) from the order in which they appear and compares the resulting "bags" of words. We used the following approach:

1. For each title in the British Library record set, downcase, eliminate stopwords, and produce a "bag of words"
2. Identify each HathiTrust record whose title matches to any of the words in the bag
3. Determine precision (P) and recall (R), calculate an average confidence score, rank by score

The resulting output is a list of candidate matches from the HathiTrust collection for each British Library record, with a corresponding confidence score. For example:

****

BL title: “St. Paul at Philippi. A Seatonian poem.”

Bag of words: st,paul,philippi,seatonian,poem

Score (P) (R) Matching HathiTrust Title

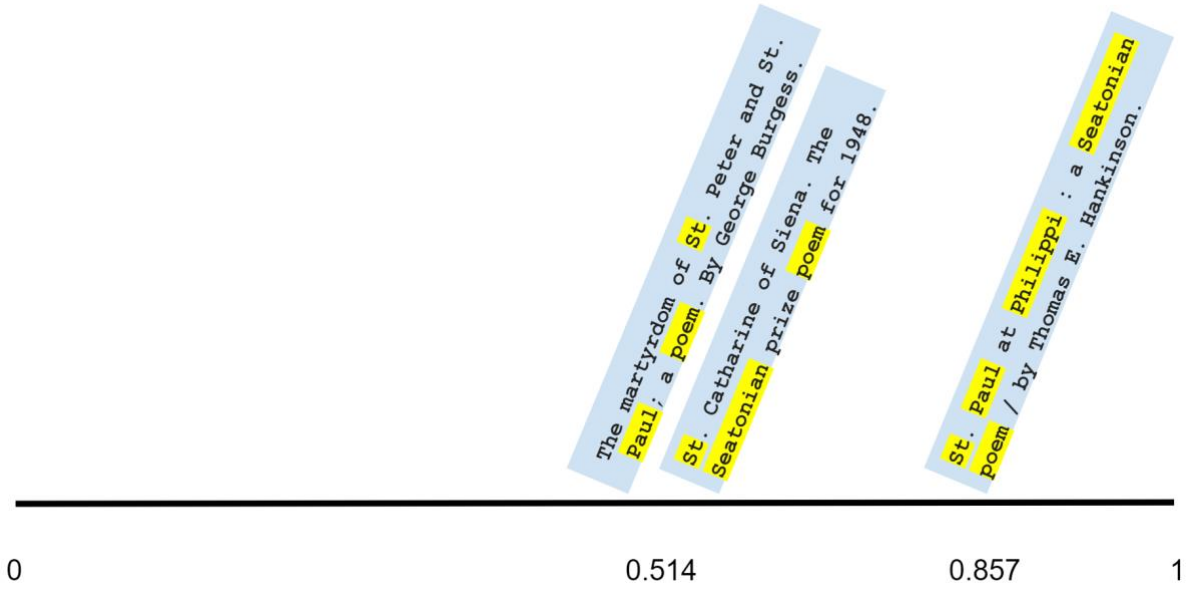
0.514 0.600 0.429 The martyrdom of St. Peter and St. Paul; a poem. By George Burgess.

0.514 0.600 0.429 St. Catharine of Siena. The Seatonian prize poem for 1948.

0.857 1.000 0.714 St. Paul at Philippi : a Seatonian poem / by Thomas E. Hankinson.

****

Confidence rankings can be expressed visually along a continuum, like so:



In aggregate, these rankings could be used to establish a “threshold” confidence score that could serve as an effective mechanism for eliminating non-matches, thereby reducing the pool of candidate matches to a more manageable size.

Machine learning

Our final approach builds on unpublished work done by Michael Morris-Pearce, formerly of the California Digital Library, to explore whether we could train an algorithm to distinguish between “good” and “bad” matches. We posed the question this way: “Given a set of distances, or differences, between fields in a pair of records, can an algorithm correctly guess whether both records in the pair refer to the same item (or not)?” Using a standard machine learning process

(setup, training/iteration, implementation) and the Python scikit-learn machine learning library, we performed two rounds of testing: one using pre-matched HathiTrust records as training data and utilising the Damerau-Levenshtein method[13] to calculate the "difference" between pairs of records; and the other using the OCN-matched NLS-HathiTrust records referenced above as training data, and utilising cosine similarity[14] to calculate difference. Results are shown below:

Round 1

Algorithm	**# of clusters in test set**	**# of correctly predicted clusters**	**Precision**	**Recall**
Polynomial	5989	5558	.982	.911
RBF SVC	5989	5948	.975	.968

Round 2

Algorithm	**# of clusters in test set**	**# of correctly predicted clusters**	**Precision**	**Recall**
Regression	83894	74029	.937	.882
Stochastic Gradient Descent	83894	71538	.928	.853
Linear SVC	83894	74025	.928	.882
RBF SVC	83894	69731	.940	.831

These methods evaluate many more dimensions than the earlier approaches we tried, so we *presume* they are better at dealing with short titles, common titles, etc. We have confidence in the approach in general, however additional training and tweaking was needed after the first round of testing. For example, recall scores are higher in the Round 1 results, but this finding seemed suspect due to the relatively small number of records in the training data. In the second round we used a bigger dataset, tested additional algorithms, and also tried an alternative method for measuring distances.

Note that we tested the Radial Basis Function (RBF) algorithm in both rounds, and both results

[13] Damerau, Fred J. (March 1964), "A technique for computer detection and correction of spelling errors", Communications of the ACM, 7 (3): 171–176, doi:10.1145/363958.363994

[14] Citation needed; check this resource: C.D. Manning, P. Raghavan and H. Schütze (2008). Introduction to Information Retrieval. Cambridge University Press.

are reported in the tables above. The polynomial algorithm used in Round 1 crashed when we applied it to the larger dataset in Round 2, suggesting that more study is needed to determine the relative computational expense of each algorithm, especially when evaluated against precision/recall results.

Insights and challenges

As expected, the UK library records contained too few OCNs for this type of identifier to serve as a reliable match point. Similarly, the presence of other identifiers such as ISBN is likely too variable to be useful for matching. Other methods show varying levels of promise: literal string matching is simple to code but very rigid; word-by-word matching is more complicated but presents a more nuanced view. Machine learning methods appear to be more accurate, but require training time and are resource-intensive to run. Note that the machine learning approaches we explored for the project are not novel, but they were new to us, and we had a unique opportunity to apply them to these particular records for evaluation purposes.

We identified a number of challenges in executing this phase of the project. For example, we encountered two situations where results were hard to interpret: first, we identified over 600 BL records containing something resembling an OCN, and second, we identified a very high number of NLW print materials containing something resembling an ISBN. Both findings were unexpected, given the age of the materials described by those records and the relatively recent introduction of the two identifier schemes. Confirming or disproving these findings will require further interrogation.

In addition, we noted that author and date fields are not well-standardized across records and therefore don't work well as string components for literal matching, yet author and date could serve as important disambiguation points for works with very similar titles. In the same way, stopword management is an effective tool for increasing the efficiency of word-by-word matching (and could have a beneficial effect on precision and recall, as well), but is hampered by language differences that are not easily overcome. Finally, both training datasets used in the machine learning work relied on OCNs for matching, however OCNs themselves are not always a reliable indicator of whether two records refer to the same work.

The team concluded that record matching is difficult, context specific work that involves trade-offs, and that there is likely no one-size-fits-all approach that will meet every need. Our overall experience suggests that a cascading approach may be best: use identifiers when they are available, then apply string matching, word-by-word matching, and finally ML methods. Such an approach would progress from simple, easy to code and less resource-intensive methods, eliminating non-matches and reducing the pool of candidate matches at each stage.

This phase of work also highlighted the importance of user interface design in any web-based presentation of the dataset. To adequately meet our use cases, the interface must be able to help users distinguish between multiple copies of the same work in meaningful ways. This has implications for clustering and faceted search implementation, and should be considered fully in a future phase.

Aggregation

Our next task was to aggregate the records from all four libraries into a single dataset. Our goal was to create an aggregation of the records from all four partners containing enough bibliographic information to be useful, while also managing the size of the resulting file. To do this, we evaluated each set of records to identify common fields and assessed the prevalence of those fields across the records. In this way, we were able to create a data file consisting of records that were reasonably complete and comparable between institutions, while also being mindful of any future

applications or hosting that might be negatively impacted if the file were too large. Even with only these limited fields, the 17.5 million records resulted in a 5 gigabyte plaintext file once aggregated together.

Data Element	**Description**	**MARC field/s**
Volume Identifier	Permanent item identifier	-
Title		245|a
Imprint	Publisher + Date of publication	260|bc
Publication Place		008 (bytes 15-17)
Author		100|abcd; 110|abcd
(URI)	Link to digital object	856
(Publication Place)		260|a

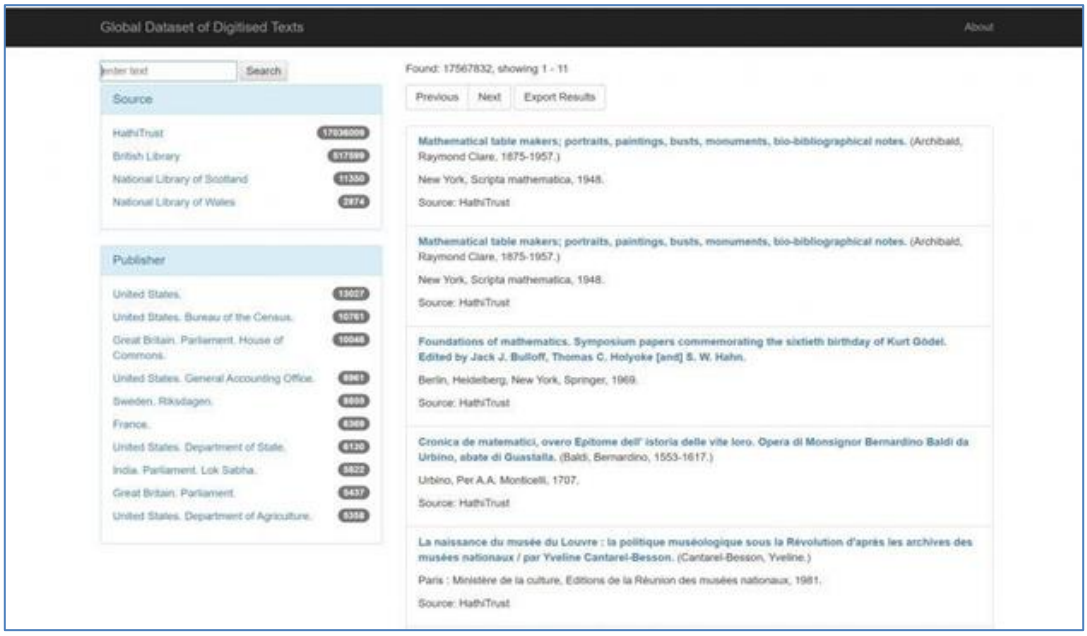
The dataset also includes the source of each record in a separate field. This information was not present in the original records; we added it as part of the aggregation.

We did not include subject headings because the HathiFile records we worked from did not contain them, and we did not include language codes because their presence across the four sets of records was too variable. The aggregation exercise also highlighted an inconsistent practice in the use of item URLs: one of the partners provided records that have URL links to the source library catalogue rather than to the item itself; other records pointed directly to the digitised items at the source library, and still others pointed to digitised copies hosted on the Google Books site.

Project Outputs

As a result of the work, we were able to produce two concrete outputs. The first of these, described above, is a 5GB CSV file containing 17.5 million records in plain text. This is an easy-to-use format which can be manipulated in widely available tools including spreadsheet software such as Excel. This CSV file has been made available under a Creative Commons By-Attribution License via the National Library of Scotland's Data Foundry.[15] It represents both a snapshot of the partner libraries' digitised collections, and an openly licensed community resource that we hope will form the basis for further analysis into the collections as data. This opens up opportunities for others to build upon the work outlined above, to undertake computational analysis of the records, and to get a small snapshot of the scale of the digitised holdings across the world's libraries.

Additionally, while we identified core use cases for a global register of digitised texts, it was evident that we also needed to interrogate the data to understand how it might meet those cases. To fulfil several use cases, for instance, it was necessary to create an easy interface for users to first search for items (either single items, or groups of items) and then to navigate to the item, or in the case of large collections of items, download the relevant URLs. The scope of the project meant that we did not have any dedicated resource to develop a custom search engine, so instead we built a lightweight prototype search interface. Led by the National Library of Scotland, the search interface utilised a SOLR index, with a faceted search interface provided by the Soldrdora system by Hector Correa.[16] This allowed us to produce a search system:



The prototype allowed us to more easily explore the dataset, and to evaluate whether the current data is sufficient to meet the core use cases. For example, some immediate issues come to light when trying to use the system: many of the items are only available to members of the particular libraries, but access statements or rights information are not displayed to inform the reader whether a particular record is accessible. Other than source (the Library that provided the records) and the publisher, there didn't appear to be other fields with enough consistency to apply useful facets on. Finally, the 'export results' button was not fully developed, something that would need to be implemented to meet use cases where a user would want to download a file containing their complete search result.

15 This is available from the NLS Data Foundry website, at the following DOI: https://doi.org/10.34812/fda4-5336.

16 Further details about the creation of the prototype can be found at the GDDNetwork website: https://gddnetwork.arts.gla.ac.uk/index.php/2019/12/23/from-aggregation-to-prototype/.

The prototype offered an opportunity to assess the dataset against the use cases, but unlike the dataset it was only ever intended to be a short-term tool. In order to more fully address our core use cases, it would be necessary to develop a more robust and sustainable discovery platform, and one focus of future work should be to develop a roadmap towards doing so. However, while the prototype is no longer available, it could be easily replicated using the method utilised by the National Library of Scotland.

Summary

This report has established the critical context for the AHRC-funded Global Digitised Dataset Network (GDDNetwork). It has also outlined the network's work in two areas: first, to establish use cases for the dataset; and second, to establish workflows to support holdings analysis and metadata matching. This work allowed us to create and release an aggregated dataset containing approximately 17.5 million records from HathiTrust, the British Library, the National Library of Scotland and the National Library of Wales. This dataset is now openly available and will provide a long-term legacy for the work of the project. We were also able to build a prototype search interface that allowed us to evaluate whether the current dataset is adequate to meet core use cases.

Our findings are:

- There still exists no global resource within the library and heritage sector that comprehensively aggregates descriptive, preservation, or provenance metadata for all digitised texts: this impacts upon discoverability of digitised materials, and limits the impact of library digitisation efforts.
- Several clear use cases emerge that would differentiate such a resource from existing platforms. There is great overlap between the use cases we identified as "discovery and access" and the existing focus of platforms including DPLA and Europeana Collections. The cases we identified in "efficiency, cost, impact value", "provenance" and research provided a clear value proposition that is not fully addressed by any existing platform.
- Within these use case categories, there is a desire for a resource that allows collaborative strategic planning to take place: areas of particular interest are in clustering similar materials, digitisation planning, and the preservation of existing materials.
- Our dataset already allows us to address several use cases, but there are clear limitations that need to be addressed by enhancements to existing metadata, and by incorporating additional data fields in future iterations.
- Holdings analysis will inevitably have to play a key role in supporting use cases across the different categories, particularly in clustering of similar or duplicate manifestations, provenance, and discovery.
- Aggregation of diverse collections metadata remains a key challenge: clustering of similar records yields better results through resource-intensive methods. Therefore, a trade-off between efficiency and accuracy might be necessary should aggregator-side metadata transformation be adopted.
- Information professionals were generally excited by the value proposition of the GDDNetwork approach, but we found that some scholars professed confusion about the uniqueness and value of the dataset. Further work is therefore required to define and explain the scope of a global registry for non-information professionals.

The idea of a single, centralised point of ground truth for digitised materials across specified geographic areas is not new. However, there are three points of difference that provide a unique value proposition to the GDDNetwork dataset. First, a truly global resource could in time provide a platform for data sharing that would allow collective decision-making, and reshape the nature of collaboration between organisations. This objective is aligned with the direction of travel for intra-national resources that prioritise the development of national-level strategic planning, as is the case with the NBK. Our use cases suggest holdings clustering and records-matching, collection holdings analysis, and collaborative preservation planning as three potential areas of value for the library community. There are also potentially secondary use cases that the dataset might support, such as collection management or deaccessioning, which we did not explore. Second, the scholarly community would benefit not only from web-scale discovery of digitised materials, but also from the availability of access to many items at once in a computationally friendly format. The dataset itself could be an important resource for understanding collections in

a trans-national context, via data-driven approaches. It would also provide scholars with a discovery mechanism for identifying texts that are suited to computational reuse. Third, the concept of a global network provides an opportunity for the library sector to address key challenges in representation and diversity in library collections. Our survey notes that more needs to be done to truly represent the global nature of the library sector, and to make discoverable resources from diverse organisations. This will necessitate incorporating other languages, scholarly traditions and forms of publication that will require further work in the areas of ingest and data matching. It will also require us to either extend beyond the US/UK context of the original funded network to reflect a diverse global community, or to refine the concept in an explicitly narrow UK.US context.

We have identified the following priorities for future work:

- Further work is required to create a business case for building a global dataset. This should include refining the use cases to identify the primary benefits and mission of the resource, as well as some assessment of the overall cost of such a service, and the potential for short and longer term institutional hosting and financial support.
- Further funding is required to expand the prototype: this should focus both upon enhancing the dataset to meet additional priority use cases, and expanding the network to include a more representative sample of global libraries.
- The US/UK constitution of our project inevitably focuses our findings upon specific regional and linguistic contexts. More therefore needs to be done to understand the issues that a global dataset would bring through diversification, including different cataloguing standards, multiple character sets, different languages, and questions of representativeness and comprehensiveness. This is both a research and a technical challenge that requires further collaboration between researchers, libraries, and content providers.

These priorities make clear that the creation of a robust and sustainable resource is a technical, research, and community development challenge. The GDDNetwork has laid a clear groundwork for future work, but we would finish the report with two calls for action. First, the network represents a coalition of the willing, but to achieve a global perspective we must expand that coalition beyond its current size and geographical focus. Second, what we have outlined in this report requires continued external investment and commitment in order to realise the potentially transformative benefits of a comprehensive global dataset of digitised texts.

References

Bilansky, A. (2016) 'Search Reading, and the Rise of the Database', *Digital Scholarship in the Humanities*. Available at: http://dsh.oxfordjournals.org/content/early/2016/05/08/llc.fqw023.

Cebr (2018) *The Heritage Sector in England and its Impact on the Economy: A Report for Historic England*. London: Cebr. Available at: https://historicengland.org.uk/content/docs/research/heritage-sector-england-impact-on-economy-2018/.

Commission of the European Communities (2005) *Communication From the Commission to the European Parliament, the Council, The European Economic and Social Committe and the Committee of the Regions: i2010: Digitial Libraries*. COM(2005) 465 Final. Brussels: Commission of the European Communities. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52005DC0465&from=EN.

Dempsey, L. (2013) 'The Emergence of the Collective Collection: Analyzing Aggregate Print Library Holdings', in Dempsey, L. et al., *Understanding the Collective Collection: Towards a System-Wide Perspective on Library Print Collections*. Dublin, Ohio: OCLC Research. Available at: https://www.oclc.org/content/dam/research/publications/library/2013/2013-09.pdf.

Dempsey, L. *et al.* (2013) *Understanding the Collective Collection: Towards a System-Wide Perspective on Library Print Collections*. Dublin, Ohio: OCLC Research. Available at: https://www.oclc.org/content/dam/research/publications/library/2013/2013-09.pdf. (

Department for Digital, Culture Media & Sport (2017) *UK Digital Strategy 2017*. Available at: https://www.gov.uk/government/publications/uk-digital-strategy/uk-digital-strategy#data---unlocking-the-power-of-data-in-the-uk-economy-and-improving-public-confidence-in-its-use.

Department for Digital, Culture Media & Sport (2018) *Culture is Digital*. London. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687519/TT_v4.pdf.

DPLA (2019) *Digital Public Library of America Strategic Roadmap, 2019-2022: Collaborating for Equitable Access to Knowledge for all*. Available at: https://dpla.wpengine.com/wp-content/uploads/2018/01/DPLA-StrategicPlan_2015-2017-Jan7.pdf.

European Commission (2008) *Now Online: 'Europeana', Europe's Digital Library*, *European Commission - European Commission*. Available at: https://ec.europa.eu/commission/presscorner/detail/en/IP_08_1747.

European Commission (2014) *Cultural Heritage: Digitisation, Online Accessibility and Digital Preservation - Report on the Implementation of Commission Recommendation 2011/711/EU*. European Commission. Available at: http://ec.europa.eu/information_society/newsroom/image/recommendation-2011-2013_progress%20report-final-clean-shared%20with%20eac-ga%20approved-22-09-2014-final_6953.pdf.

European Union (2017) *New European Interoperability Framework: Promoting Seamless Services and Data Flows for European Public Administrations*. Luxembourg: European Union. Available at: https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf.

Europeana (2013) *Moving to new Europeana Data Model*, *Europeana Pro*. Available at: https://pro.europeana.eu/post/moving-to-new-europeana-data-model.

Europeana (2015) 'The Europeana Data Model for Cultural Heritage'. Available at: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Factsheet.pdf.

Findlay, P. (2019) 'Collection Management and Digitised Collection: Presentation to the GDD Network Workshop'. London.

Greenhall, M. (2019) *Digital Scholarship and the Role of The Research Library: The Results of the RLUK Digital Scholarship Survey*. RLUK, p. 78. Available at: https://www.rluk.ac.uk/wp-content/uploads/2019/07/RLUK-Digital-Scholarship-report-July-2019.pdf.

Hunter, D. and Brown, K. (2010) *Thriving or Surviving: National Library of Scotland in 2030*. National Library of Scotland. Available at: http://www.nls.uk/media/808985/future-national-libraries.pdf.

Jisc (2019) *National bibliographic knowledgebase (NBK)*, *Jisc*. Available at: https://www.jisc.ac.uk/national-bibliographic-knowledgebase.

Kenny, E. (2017) *History*, *Europeana Pro*. Available at: https://pro.europeana.eu/our-mission/history.

Niggemann, E., De Decker, J. and Levy, M. (2011) *The New Renaissance: Report of the 'Comité Des Sages,' Reflection Group on Bringing Europe's Cultural Heritage Online*. Brussels. Available at: http://www.eurosfaire.prd.fr/7pc/doc/1302102400_kk7911109enc_002.pdf.

Schlosser, M. (2012) 'Welcome to Digital Scholarship @ The Libraries', *Digital Scholarship @ The Libraries*. Available at: https://library.osu.edu/site/digitalscholarship/2012/12/12/welcome-to-digital-scholarship-the-libraries/.

Showers, B. (2014) *A National Monograph Strategy Roadmap*. London: Jisc. Available at: https://www.jisc.ac.uk/reports/a-national-monograph-strategy-roadmap.

Thylstrup, N. B. (2019) 'Sovereign Soul Searching: The Politics of Europeana', in *The Politics of Mass Digitization*. Cambridge, MA: MIT Press.

Whitelaw, M. (2015) 'Generous Interfaces for Digital Cultural Collections', *Digital Humanities Quarterly*, 9(1). Available at: http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html.

WorldCat (2020) *Inside WorldCat*, *OCLC*. Available at: https://www.oclc.org/en/worldcat/inside-worldcat.html.