LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 31, (2021) 1–13 | e-ISSN: 2213-056X

Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry

Sarah Ames

National Library of Scotland, Edinburgh, United Kingdom
sarah.ames@nls.uk, orcid.org/0000-0002-0118-189X

Abstract

'Collections as data' has become a core activity for libraries in recent years: it is important that we make collections available in machine-readable formats to enable and encourage computational research. However, while this is a necessary output, discussion around the processes and workflows required to turn collections into data, and to make collections data available openly, are just as valuable. With libraries increasingly becoming producers of their own collections – presenting data from digitisation and digital production tools as part of datasets, for example – and making collections available at scale through mass-digitisation programmes, the trustworthiness of our processes comes into question. In a world of big data, often of unclear origins, how can libraries be transparent about the ways in which collections are turned into data, how do we ensure that biases in our collections are recognised and not amplified, and how do we make these datasets available openly for reuse? This paper presents a case study of work underway at the National Library of Scotland to present collections as data in an open and transparent way – from establishing a new Digital Scholarship Service, to workflows and online presentation of datasets. It considers the changes to existing processes needed to produce the Data Foundry, the National Library of Scotland's open data delivery platform, and explores the practical challenges of presenting collections as data online in an open, transparent and coherent manner.

Keywords: digital scholarship; collections as data; transparency; provenance; datasets; digital humanities

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.18352/lq.10371

1. Introduction

In 2017, Thomas Padilla wrote of a 'collections as data imperative' for libraries and cultural heritage organisations, which focused on three key concepts: generativity, legibility and creativity (Padilla, 2017, p. 2). As part of this, he explains that, 'To make collections as data usable, the processes by which they are established must be made legible' (p. 3): data provenance and transparency is essential to releasing useful and usable collections as data. Padilla points out that decision making processes in libraries and the transformations that collections and data undergo are not traditionally made available to users – information which often determines why certain collections are made available above others, influencing both research agendas and information available to the public.

Two years on from this paper, the National Library of Scotland launched its Digital Scholarship Service amidst the burgeoning interest in the concept of 'collections of data' stemming from the work of Padilla and colleagues on the *Always Already Computational* project, which ran from 2016–2018 (Padilla et al., 2019a), and its Mellon-funded continuation, *Collections as Data: Part to Whole* (Padilla et al., 2019b), and started to release collections in machine-readable format. This work is part of a broader shift currently being seen in Galleries, Libraries, Archives and Museums (GLAMs) towards releasing collections openly and in machine-readable format, and to support users wishing to access or analyse these collections using computational methods. The International GLAM Labs Community[1], for example, advocates for 'laboratory'-style innovative experimentation with digital collections, and Digital Scholarship teams, services and roles are now familiar in US organisations, and increasingly becoming a core part of European research libraries (such as the British Library) and beyond.

Within this broader context of 'collections as data' activity, this case study, based on a presentation delivered at the LIBER2020 conference (Ames, 2020a), explores how the Library's open data-delivery platform, the Data Foundry, has been designed to include data provenance; how the Library's Digital Scholarship Service works to embed transparency into the Library's processes; and the practical implications, benefits and challenges of this activity. While material provenance has always been an important topic for libraries and field of book history, this paper explores why this documentation and

transparency is relevant to current library practices, particularly around digitisation and data release, and how steps can be taken to make this information available to users. It recognises that libraries are increasingly becoming data producers – and a producer of their own collections – and that this is problematic, and explores the ways in which the National Library of Scotland is approaching this issue.

2. Launching a Digital Scholarship Service

Much of the dialogue around digital scholarship and digital humanities in libraries has advocated for this as an area which is *not* a service: Muñoz (2012) put it plainly that 'Digital humanities in the library isn't a service', and more recent reports have championed the role of libraries as partner in digital humanities projects: LIBER has advocated strongly for the library as a partner (Wilms et al., 2019) and the work of RLUK (Research Libraries UK) is around libraries as 'provider, partner, pioneer' (Greenhall, 2019, p. 5). The benefits to both researchers and libraries from collaboration and partnership in this area are clear: both gain from the expertise of the other. In this context, why, then, launch a digital scholarship 'service'?

Established in September 2019, the National Library of Scotland's Digital Scholarship Service has five main objectives:

- Encourage, enable and support the use of computational research methods with the collections;
- Ensure that the collections are being used to their full potential;
- Establish a culture in the Library which supports digital scholarship;
- Practise and promote transparency in our data creation processes;
- Anticipate the future of research. (Ames, 2020b)

This case study focuses largely on the fourth of these goals, yet the aims more generally centre around three strands of activity: making the Library's collections available as data; external engagement and collaboration; and internal engagement within the Library. As a result, the Library can offer three digital scholarship service levels to its users:

1. Self-service (making use of published data collections or tools);

2. Self-service of collections or tools with some staff consultation time (for example, contributing to a class or providing collections expertise);
3. Partnership in funded projects.

Combined, these cover users who require little interaction with the Library to carry out their required task, and those who wish to collaborate on bigger projects.

A number of reasons have led to this development of a 'service', including pragmatism: 'service' is a term that is understood within libraries, enabling systematic and coordinated service development and service level agreements between teams. 'Service', while a problematic term, is also an ongoing activity: associating finite terms such as 'project' or 'programme' with this work would not accurately convey the need for this activity to become business-as-usual. Furthermore, libraries need some level of service provision to enable collaboration in the first place: offering collections in machine-readable format, or technical support, involves setting up a service and involving multiple teams to deliver these. And lastly, some users simply do not want or need to collaborate, but to make use of the Library's openly licensed data and return to working on their own project. Meanwhile, these building blocks enable collaboration as part of a funded service, which enables the Library to partner on digital scholarship projects as well.

Launching this service has involved Library-wide activity to make collections available as data. Initial activity focused on user research: based on conversations with digital humanities researchers and an assessment of user needs, as well as early usability testing of the Library's new open data platform, the Data Foundry, three broad users were identified, ranging from 'beginner' to 'advanced' – at one end of the spectrum, wanting to use online tools with text files, and at the other, with advanced technical skills, not minding what format digitised material was presented in (but having an awareness of standards and best practice). In the middle, however, was a more complex user, who had limited technical skills, but who understood the value of different formats and approaches for research questions. This user was more likely to employ a research assistant to carry out any technical work, but wanted to access data easily and quickly to find out what it was and if it was suitable for their work. Deciding to cater to this middle-ground user enabled us to satisfy many of both the beginner and advanced users also.

To ensure we could provide consistent formats from our digitisation programme, we then needed to change our approach to digitisation. Digital scholarship provides a new use case for digitisation. Where, previously, the Library had been digitising solely for online image galleries or user orders, it now needed to account for a new audience. Enabling computational uses of collections requires some additional steps, including granular rights assessment to ensure the collection can be released under the definition of open data; different file formats created (Optical Character Recognition (OCR) outputs as ALTOXML, or 'Analysed Layout and Text Object Extensible Markup Language', which provides layout information of the page); metadata to describe the digital object (METS – 'Metadata Encoding and Transmission Standard' (Library of Congress, 2020)); storage solutions suited to large downloads (the Data Foundry uses Amazon Web Services buckets); and persistent identifiers to enable citation (Digital Object Identifiers (DOIs)). Other decisions needed to be made about consistent standards, image sizes, which datasets to select to make available first, and what additional information about these processes that it would be possible to gather and make available – how to be transparent and to present the Library's datasets in context. Following this activity, in September 2019 the Data Foundry, the data-delivery platform for the Digital Scholarship Service, was published and launched online (National Library of Scotland, 2019a).

3. Embedding Transparency in Library Practices

Amidst this work, how is the Library promoting transparency in line with the Digital Scholarship Service's fourth objective, and what are the challenges of this? Furthermore, why be transparent – what is the value of transparency? Open practices enable the Library to convey information about internal processes, any transformations the collections or data have undergone, and the decision making around this. This will also enable libraries to acknowledge and reduce biases – to be open about where decision making processes may not be perfect, where they may have failed, and where historical decision making or practises still have repercussions today. And this empowers libraries and their users to provide counter-narratives: to explore what is missing, what hasn't been said, what gaps there are, and to start to rectify this.

To date, the Digital Scholarship Service has focused on three main ways of promoting and embedding transparency in Library practise and processes.

3.1. Communicating Transparency

With the Library's remit spanning a broad audience – not solely the research community – one straightforward way to communicate with a large number of users has been through social media. One example of this was the Library's Twitter activity for the Day of Digital Humanities 2020 (#DayOfDH2020), where the Digital Scholarship Service highlighted 'invisible labour' and digital scholarship. As Poster et al. (2016) explain, invisible labour can include 'visible work done by invisible people (domestic workers, librarians)' (p. 7), Graban et al. (2019) write that, 'Much of the work that takes place in DH projects is invisible to multiple levels of authority'. The end-product of digitisation processes – the availability of rights-assessed, digitised material online – is a prime example of this work, with the effort to digitise a single item spanning multiple library teams and processes.

Highlighting this involved a Twitter 'thread' explaining the processes a collection undergoes when being turned into data, to lift the lid on the decision making and transformations, and encourage an understanding of the amount of work and the number of teams involved in this (National Library of Scotland, 2020). Taking care to word the thread for a broad audience, steps including curatorial work, rights assessment, conservation, bookfetching, digitisation (capture), ingest and preservation were all included, as was information about the labour involved from metadata and developer teams; details of what a 'dataset' is, and highlighting the importance of transparency in these processes; the storage location of digital objects; and the role of persistent identifiers.

This enabled the Library to show users 'under the hood' of one of the areas of activity of the Digital Scholarship Service, in the comfort of their own homes, enabling us to highlight the human effort – and therefore biases – being put into our collections. It also served to begin a dialogue with the Library's users about the implications of this work for the collections that are released and what they can be used for. A similar Twitter 'thread' was published to explain Handwritten Text Recognition (HTR) technology to a public audience, to

mark the release of the Library's first artificial intelligence (AI)-generated dataset using the Transkribus platform (National Library of Scotland, 2021), enabling the Library to communicate the human involvement in AI-generated work, and the problems with this.

3.2. Value of Clarity in Online Presentation of Data

What do all of these processes mean for the way in which we convey data to users? Amidst the multiple teams and transformations that an item passes through as it becomes a 'dataset', the presentation of data itself, and the way it is communicated to audiences, becomes important. The National Library of Scotland has an Open Data Publication Plan, which sets out the level of open data the Library works to (three star), which formats are used, how frequently data is published and where it is published (National Library of Scotland, 2019b). Complementing this is an Open Data Register of all of the datasets the Library has published, their dates, update frequency and file formats; both can be downloaded from the Data Foundry (National Library of Scotland, 2019c). This enables the Library to set out clear expectations around when and how data is published.

Datasets themselves are published to the Data Foundry website. The presentation of cultural heritage datasets is extremely important: given their size and the need to download the data to begin to explore it, the concept of a collections dataset can appear to be somewhat opaque. Padilla (2017) notes that 'the quality of data [is not] typically indicated' (p. 3), and this, alongside information about the size of the data – conveyed in both bytes and words – becomes important for understanding cultural heritage datasets. By issuing a number of key points about each dataset, such as the number of files; number of words and lines; level of OCR clean-up; and date range of the data – essentially, declaring the corpus – as well as providing a curator-led introduction to the collection, datasets on the Data Foundry begin to provide some shape for an otherwise abstract concept (Figure 1).

Furthermore, all of the Library's digitised collections go through a prior rights assessment process, so all datasets on the Data Foundry have clear licences or rights statements, and the Library does not assert further copyright over collections. Rights statements or licences are added to each dataset;

Fig. 1: Example of a dataset on the Data Foundry, with key points about the data provided.

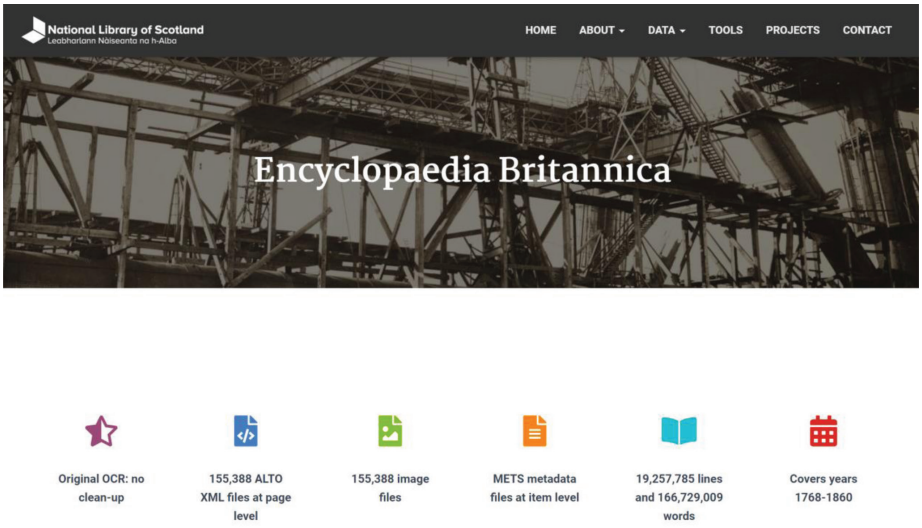


Fig. 2: A dataset spanning two rights statements.

some collections, such as Encyclopaedia Britannica, which span lengthy time-periods, cross into a number of rights statements (Figure 2).

Data Foundry datasets are provided as simple downloads to reduce technical barriers to use, with future plans for Application Programming Interface (API) access to some collections. Each dataset includes an inventory file and readme file, which includes high-level information about the dataset, including file numbers and formats, date of publication and subsequent revisions, and rights information. Including METS, ALTO and image files in

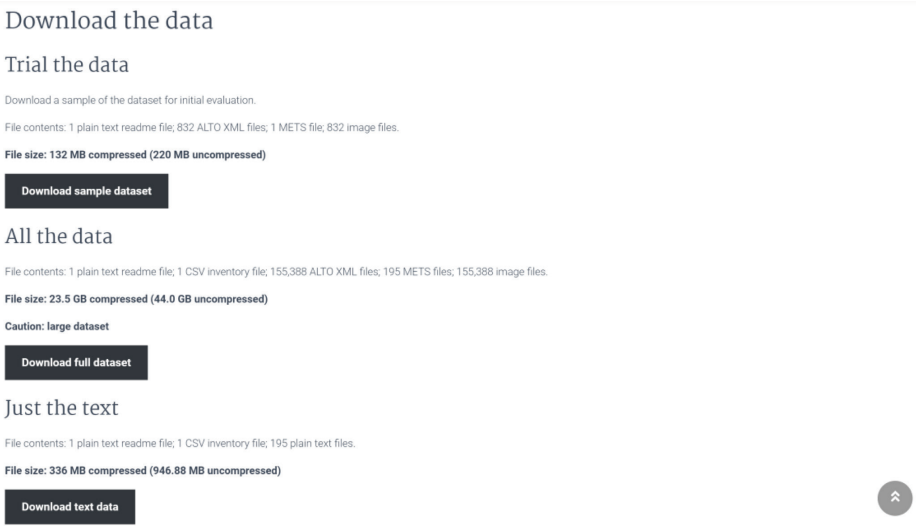
these packages led to some datasets reaching more than 40GB in size when uncompressed, so storing and making them available in the cloud ensures fast downloads. The Data Foundry provides tiered downloads of the data, to enable users to download a sample before committing to downloading the whole dataset, and also to enable users who only wish to access plain text files (for example, in the case of the 'beginner' user referred to above) (Figure 3).

When the dataset is compiled, zipped and published online, a DOI is then added as a persistent identifier to enable citation. Together, the features and files of the Data Foundry are designed to ensure minimal effort is needed to gain an overview of the data before committing to downloading it.

3.3. Data Provenance

Finally, conveying the provenance of cultural heritage data is an essential of retaining the thread from a data collection back to a print collection. Ensuring that readers understand where the data they use has come from, what processes it has undergone and why it exists in the first place is important for a

Fig. 3: Tiered dataset downloads on the Data Foundry website.



number of reasons; yet how to convey that information in an area which currently has few agreed standards is a challenge.

The use of METS enables the Library to include provenance information such as descriptive metadata relating to bibliographic and holdings information; administrative data, including information about image capture, file creation and rights; file listings; and structural metadata. This kind of information, and particularly the technical provenance information detailing the digitisation process, is essential to be able to reconstruct the actual print process from print collection to data collection, aligning libraries with the reproducible research movement (Ames & Lewis, 2020; Padilla, 2017). However, this does not extend as far as information about *why* an item or collection was digitised in the first place: as Padilla (2017) points out, 'Libraries do not typically expose why some collections have been made available and others have not' (p. 3).

Why is this important? As Ames and Lewis (2020, p. 4) point out, a library's collection as a whole is a problematic concept – often informed by biases in historic (and current) collecting practices. Adding to this, only some items from this collection are digitised, and this depends on factors including copyright, conservation and internal selection processes. From these digitised collections, only some are then suitable to make available as datasets – depending on granularity of rights assessment to make collections available as 'open' data, OCR quality or prior digital asset management practices. 'How to present these collections in context, and how these thinned-down collections could become representative of a broader, tacit understanding of 'culture', is problematic' (Ames & Lewis, 2020, p. 4) – and directly influences the resources available for research fields, creative uses and school learning. Furthermore, how these collections are representative of the diverse communities that libraries service is often problematic – as is how we explain this problem of the skewed availability of digital material to library users.

As a result, datasets on the Data Foundry include information about the reasoning behind how and why datasets have been produced, to ensure a fuller understanding of the collections: this is an important aspect of contextualising libraries' digital and digitised collections. The Library is working towards systematically making available information about why certain collections have been digitised, what funded them, whether they were outsourced for

digitisation and what curatorial decisions were made when proposing the collection for digitisation. This enables the Library to present information about why the collection came about in its current digital form, and also enables us to build freedom of information into our data creation processes: being open about this kind of information can create additional efficiencies. With no current standards for how to provide this information, it is currently provided in an unstructured, free-text 'Other' field in METS files for datasets, as a method of keeping this information directly alongside the datasets.

4. Responsibilities of Libraries: Conclusions

All of this activity points towards the changing responsibilities of libraries. Traditionally, libraries have been acquirers: they have purchased items or collections, or they have been donated, and they have then described them according to (supposedly) objective, neutral standards – which is problematic in itself – and then made them available to users for borrowing or research. With the shift towards digital, however, we are now seeing libraries increasingly becoming producers of their own collections: at the National Library of Scotland, for example, data is not only produced as digitised texts are OCRed, but there is also data created by the photography equipment, the OCR software itself, the digital production tools used for post-processing. Like many libraries, we also produce metadata, and we produce organisational 'corporate' data.

What, then, are libraries' responsibilities as data producers? How do libraries describe their own data and convey information about how and why it has been produced? How can they – and can they ever – accurately contextualise their own processes? And how do they remain in a position of trust, amidst a backdrop of 'fake news', artificial intelligence and Deepfakes, when they are creating that information themselves? Transparency becomes essential: the more that libraries can set data collections in context, and be clear about the processes and decision making they have undergone, the more value they have as research material. This activity remains work in progress, yet work around the 'collections as data imperative' by libraries becomes more important by the day as we seek to ensure that library collections remain relevant, usable and useful as technologies, methodologies – and the broader world around us – change.

References

Ames, S. (2020a). *Transparency, provenance and collections as data: the National Library of Scotland's Data Foundry*. [Conference presentation]. 50th LIBER Annual Conference – LIBER2020, online. http://doi.org/10.5281/zenodo.3921784.

Ames, S. (2020b). *Digital scholarship and the Data Foundry*. 21st Century Curator talk series, British Library. http://doi.org/10.5281/zenodo.3862050.

Ames, S., & Lewis, S. (2020). Disrupting the library: Digital scholarship and big data at the National Library of Scotland. *Big Data & Society*, *7*(2), 1–7. https://doi.org/10.1177/2053951720970576.

Graban, T. S., Marty, P., Romano, A., & Vandegrift, M. (2019). Questioning collaboration, labor, and visibility in Digital Humanities research. *Digital Humanities Quarterly*, *13*(2), 1–9. http://www.digitalhumanities.org/dhq/vol/13/2/000416/000416.html.

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

Library of Congress. (2020). *METS - Metadata encoding & transmission standard - Official website.* https://www.loc.gov/standards/mets/.

Muñoz, T. (2012, August 19). *Digital humanities in the library isn't a Servic*e. [blog] http://trevormunoz.com/notebook/2012/08/19/doing-dh-in-the-library.html.

National Library of Scotland. (2019a, September 2). *Data Foundry launched*. https://www.nls.uk/news/archive/2019/09/data-foundry.

National Library of Scotland. (2019b, September 2). *Open Data publication plan*. https://data.nls.uk/download/national-library-of-scotland-open-data-publication-plan.pdf.

National Library of Scotland. (2019c, September). *Standards-Open Data Plan*. https://data.nls.uk/about/standards/.

National Library of Scotland. (2020, April 29). *Data foundry.* Twitter. https://twitter.com/natlibscot/status/1255478042207621126.

National Library of Scotland. (2021, January 28). *Handwritten Text Recognition (HTR).* Twitter thread. https://twitter.com/natlibscot/status/1354751444004597762.

Padilla, T. (2017, February 15). *On a Collections as data imperative*. Library of Congress. https://labs.loc.gov/static/labs/work/reports/tpadilla_OnaCollectionsasDataImperative_final.pdf.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019a). *Final report – Always already computational: Collections as data.* Zenodo. https://doi.org/10.5281/zenodo.3152935.

Padilla, T., Scates Kettler, H., Varner, S., & Shorish, Y. (2019b). *Collections as data: Part to whole*. Github. https://collectionsasdata.github.io/part2whole/.

Poster, W. R., Crain, M., & Cherry, M. A. (2016). Introduction: Conceptualizing invisible labor. In M. Crain, W. R. Poster, & M. A. Cherry (Eds.), *Invisible labor: Hidden work in the contemporary world* (pp. 3–27). University of California Press. https://doi.org/10.1525/9780520961630-003.

Wilms, L., Derven, C., O'Dwyer, L., Lingstadt, K., Verbeke, D., & Lefferts, M. (2019). *Europe's Digital Humanities landscape: A study from LIBER's Digital Humanities & Digital Cultural Heritage Working Group*. Zenodo. http://doi.org/10.5281/zenodo.3247286.

Note

[1] International GLAM Labs Community. https://glamlabs.io/.