

Building the Picture Behind a Dataset

Frances Madden
The British Library

Jez Cope
The British Library

Abstract

As part of the European Commission funded FREYA project The British Library wanted to explore the possibility of developing provenance information in datasets derived from the British Library's collections, the data.bl.uk collection. Provenance information is defined in this context as 'information relating to the origin, source and curation of the datasets'. Provenance information is also identified within the FAIR principles as an important aspect of being able to reuse and understand research datasets. According to the FAIR principles, the aim is to understand how to cite and acknowledge the dataset as well as understanding how the dataset was created and has been processed. There is also reference to the importance of this metadata being machine readable. By enhancing the metadata of these datasets with additional persistent identifiers and metadata a fuller picture of the datasets and their content could be understood. This also adds to the veracity and understanding the dataset by end users of data.bl.uk.

Submitted 16 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Frances Madden, The British Library, 96 Euston Road, London, NW1 2DB, United Kingdom. Email: frances.madden@bl.uk

This paper was presented at International Digital Curation Conference IDCC20, Dublin, 17-19 February 2020

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: <http://www.ijdc.net/>

Copyright rests with the authors. This work is released under a Creative Commons Attribution Licence, version 4.0. For details please see <https://creativecommons.org/licenses/by/4.0/>



Introduction

As part of the European Commission funded FREYA project, the British Library wanted to explore the possibility of developing provenance information in datasets derived from the British Library's collections, the data.bl.uk collection. Provenance information is defined in this context as 'information relating to the origin, source and curation of the datasets' (Lavasa et al., 2019). Provenance information is also identified within the FAIR principles as an important aspect of being able to reuse and understand research datasets.¹ According to the FAIR principles, the aim is to understand how to cite and acknowledge the dataset as well as understanding how the dataset was created and has been processed. There is also reference to the importance of this metadata being machine readable. By enhancing the metadata of these datasets with additional persistent identifiers and metadata a fuller picture of the datasets and their content could be understood. This also adds to the veracity and understanding the dataset by end users of data.bl.uk.

Background

data.bl.uk

data.bl.uk is the Library's collection of datasets derived from its collections, mostly created through the British Library Labs project. These are held in a wide variety of formats including images, xml, csv and three-dimensional models. To date these are made accessible via a section of the Library's website and have DataCite DOIs assigned to them on deposit in the system, however the limits of this platform are well understood, despite continued usage since 2016, with over 700 resolutions to the top 10 DOIs with the BL Labs prefix in the first six months of 2019. The Library is undertaking a pilot repository project to develop a shared repository with other UK Independent Research Organisations including the British Museum, National Museums Scotland, Tate Galleries, Museum of London Archaeology and Royal Botanic Gardens Kew. The Library partnered with Ubiquity Press to develop the repository based on Samvera Hyku. data.bl.uk was one of the collections targeted for immediate migration during the repository's development so as part of the development, the data.bl.uk collection was loaded to the Hyku platform, where the functionality was tested and went live as a beta in November 2019. The repository provides enhanced functionality including integrated DOI minting capabilities for DataCite DOIs, a customisable metadata model, improved search functionality and download capacity for large data files.

FREYA context

The FREYA project aims to extend the infrastructure around persistent identifiers through creating a PID Graph which connects research entities together using persistent identifiers.² The British Library is one of the partners in FREYA, representing the arts, humanities and social sciences. It is one of twelve partners, representing a mix of disciplines, persistent identifier service providers and publishers. One of the key contributions by the British Library in the context of this project is to develop enhanced provenance information relating to datasets and theses held in EThOS, the index of UK theses maintained by the Library.

¹ <https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/>

² <https://www.project-freya.eu/en/pid-graph/the-pid-graph>

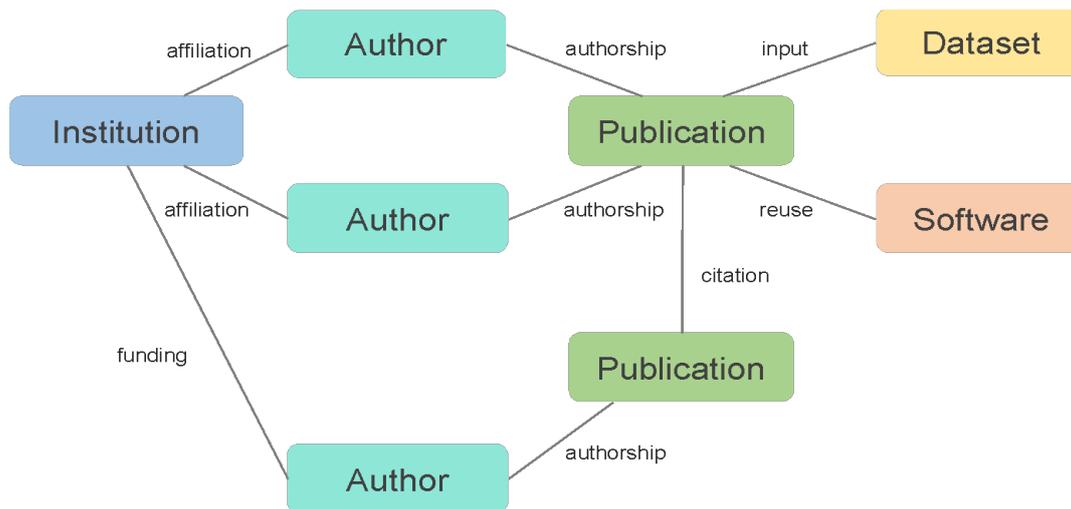


Figure 1. An example PID Graph

The major output of FREYA is the PID Graph, see Figure 1, which connects different types of entities related to research together using PIDs. This graph includes organisational IDs for institutions, funders and research facilities, author IDs (including ORCID and International Standard Name Identifier (ISNI)) and IDs for outputs such as DOIs for papers, datasets and software. The aim is for data.bl.uk to be a PID Graph where all entities related to the dataset record have PIDs. This provenance work focuses on building these connections between the entities supporting the dataset to provide a detailed picture which would help researchers understand the dataset more comprehensively.

The specific developments identified by FREYA were informed by an exercise conducted in mid-2018 which involved gathering user stories. These user stories were collected in the standard format: ‘As a <ROLE>, I would like <FUNCTIONALITY> so that <BENEFIT>.’ A wide range of stakeholders were consulted including the project partners, their networks, FREYA’s ambassadors, and attendees of several workshops including DI4R in October 2019.

This type of metadata enhancement has the potential to build a fuller understanding of the contents of digitised collections through metadata. There will also be a possibility through further developments to capture comprehensive information about the frequency of terms and entities within collections.

Methodology

In order to create this graph, a couple of pilot datasets whose metadata could be enhanced were identified and augmented with machine readable metadata. One of these was a collection, *Theatrical Playbills from Britain and Ireland*³, a series of bound volumes of playbills from theatres across the UK and Ireland from the mid 18th century to the mid 19th century, which had been digitised. The digitised files are available along with metadata in a csv file as supporting documentation. From this it was possible to identify the ARK identifiers which were assigned to the individual playbills within the dataset and also the theatres for which the playbills were created.

186 ARK identifiers for the playbills were added to the repository record as related identifiers with the relation Has Part, see Figure 2. Due to the current display in the repository, these links are not currently actionable but it is hoped that these will be improved with the second stage of development of the repository.

³ <https://doi.org/10.21250/pb1>

From *Theatrical Playbills from Britain and Ireland* Dataset

Related identifier	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588691.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588707.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588883.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588879.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588891.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588887.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588893.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588901.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022589010.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588895.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588923.0x000002 type: ARK relation: Has Part
	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588711.0x000002 type: ARK relation: Has Part

Figure 2. A screenshot of the record *Theatrical Playbills from Britain and Ireland*. The ARK Identifiers which were added as Related Identifiers to the record.

To create a relationship between the dataset and other entities related to it, International Standard Name Identifiers, (ISNIs) for the theatres were found and created by the team at the Library where none existed in line with the dates of the theatres. It was intended that these also would be included as related identifiers. However, the repository's metadata schema, which is based on the DataCite metadata schema v.4.1, defines related identifiers as identifiers of related resources with a mandatory relation type of e.g. Cites; References or IsVariantFormOf (DataCite Metadata Working Group, 2017). As an alternative, the theatre names and ISNIs were included as Contributors, type: Other to the dataset, see Figure 3. These displayed as actionable linked icons on the page.

From *Theatrical Playbills from Britain and Ireland* Dataset

METADATA

Resource type	Dataset
Collections	British Library Datasets
Contributors	Lyceum Theatre (London, England) (Other)  Princess's Theatre (London, England) (Other)  Theatre Royal (Bath, England) (Other)  Theatre Royal (Birmingham, England) (Other)  Theatre Royal (Bristol, England) (Other)  Theatre Royal (Dublin, Ireland) (Other)  Theatre Royal (Liverpool, England) (Other)  Theatre Royal (Manchester, England) (Other)  Theatre Royal (York, England) (Other)  Olympic Theatre (London, England) (Other)  Theatre Royal (Hull, England) (Other)  Theatre Royal (King's Lynn, England) (Other)  Theatre Royal (Edinburgh, Scotland) (Other)  Theatre Royal (Stafford, England) (Other)  Theatre Royal (Margate, England) (Other)  Theatre Royal (Scarborough, England) (Other)  Theatre Royal (Portsmouth, England) (Other)  New Theatre Royal (Hull, England) (Other)  Drayton Theatre (Market Drayton) (Other)  Drury Lane Theatre (London, England) (Other)  Covent Garden Theatre (Other)  Haymarket Theatre (London, England) (Other)  Old Vic Theatre (London, England) (Other)  (Other) 
Institution	British Library
Publisher	British Library
Place of publication	London, UK
Official URL	https://doi.org/10.21250/pb1
Related URL	https://doi.org/10.21250/pb2
Licence	CC Public Domain Mark 1.0

Figure 3. A screenshot of the record *Theatrical Playbills from Britain and Ireland* illustrating the addition of ISNIs to the theatre names as actionable links through the ISNI icon.

The other pilot dataset relates to the metadata held in the EThOS index of UK theses⁴. Current institution is a controlled field within the EThOS database which relates the thesis to a particular UK higher education institution. This controlled list of 143 institutions was matched against the ISNI database and the institutions and their ISNIs were added as Contributor, type Other to the dataset, see Figure 4. Again, related identifier would have been the preferred

⁴ <https://doi.org/10.22021/ETHOSCSV201803>

location for this metadata but Contributor was a reasonable location for the metadata in this instance. The addition of the small number of DOIs which are held within EThOS to the record was explored but not actioned due to the current limitations of display.

From *UK Doctoral Thesis Metadata from EThOS* Dataset

Contributors	
	Abertay University (Other) 
	Aberystwyth University (Other) 
	Anglia Ruskin University (Other) 
	Aston University (Other) 
	Bangor University (Other) 
	Bath Spa University (Other) 
	Birkbeck (University of London) (Other) 
	Birmingham City University (Other) 
	Bournemouth University (Other) 
	Brunel University (Other) 
	Bucks New University (Other) 
	Canterbury Christ Church University (Other) 
	Cardiff Metropolitan University (Other) 
	Cardiff University (Other) 
	City, University of London (Other) 
	Coventry University (Other) 
	Cranfield University (Other) 
	De Montfort University (Other) 
	Durham University (Other) 
	Edge Hill University (Other) 
	Edinburgh Napier University (Other) 
	Glasgow Caledonian University (Other) 
	Glasgow School of Art (Other) 
	Glyndwr University (Other) 
	Goldsmiths College (University of London) (Other) 
	Harper Adams University College (Other) 
	Heriot-Watt University (Other) 
	Imperial College London (Other) 
	Institute of Advanced Legal Studies (Other) 
	Institute of Cancer Research (University Of London) (Other) 
	Institute of Classical Studies (Other) 
	Institute of Commonwealth Studies (Other) 
	Institute of Historical Research (University of London) (Other) 
	Keele University (Other) 
	King's College London (University of London) (Other) 
	Kingston University (Other) 
	Lambeth Palace Library (Other) 
	Lancaster University (Other) 
	Leeds Beckett University (Other) 
	Liverpool Hope University (Other) 
	Liverpool John Moores University (Other) 
	London Business School (University of London) (Other) 

Figure 4. A screenshot of the record *UK Doctoral Thesis Metadata from EThOS* illustrating the addition of ISNIs to the institution names as actionable links through the ISNI icon.

Findings

This piece of work was an initial pilot and the aim was to establish practices and feasibility for this going forward. One finding was the number of identifiers which were captured. This work was undertaken manually but the time taken to manually populate 100+ identifiers into an individual record demonstrated a need to use bulk uploading processes where possible. The repository development partner, Ubiquity Press have provided a bulk uploader so it is possible to do this in a less manual fashion and ensure higher quality assurance.

The volume of identifiers has also demonstrated the need for a flexible display within the repository. The current instance of the repository does not have any capability of collapsing the display or using pop-ups to provide further information. In the original repository interface, the download links to the digital objects themselves were displayed at the bottom of the record compounding the issue. This was improved by moving the links to the top of the page below the record description so it appears without the need for scrolling on a desktop computer.

The requirements of the DataCite metadata schema also highlighted the novelty of this use case for PID metadata. While it was possible to express a relationship between the theatres and the dataset and the universities and the dataset, it was not considered to be entirely accurate. For the playbills dataset particularly, the theatres did not play an active contributory role to the dataset itself. In addition ContributorType is a defined list in the DataCite Metadata Schema v4.1 comprising:

- ContactPerson
- DataCollector
- DataCurator
- DataManager
- DistributorEditor
- HostingInstitution
- Producer
- ProjectLeader
- ProjectManager
- ProjectMember
- RegistrationAgency
- RegistrationAuthority
- RelatedPerson
- Researcher
- ResearchGroup
- RightsHolder
- Sponsor
- Supervisor
- WorkPackageLeader
- Other (DataCite Metadata Working Group, 2019, p.21)

None of the options on this list (except 'Other') represented an accurate description of the relationship. As stated above Related Identifiers are intended for related resources rather than related persons or organisations. A type of related identifier such as related organisation or individual or even, more generically, related entity with which an identifier could be associated would describe the relationship more accurately.

The current schema within the repository supports ISNIs for individual and organisational contributors and creators. However there are a number of PIDs which could be assigned to organisations including the General Research ID (GRID), the Research Organisations Registry (ROR) and Fundref IDs for funding institutions. It is not clear which identifier from these would be the most suitable for the repository to use and if that would vary across different contexts. There are plans to introduce support for ROR IDs into the repository in early 2020. It is possible all could be accommodated with the schema however this might present usability issues for end users who could struggle to navigate a record with metadata elements which are

assigned multiple identifiers each e.g. Contributors with an ISNI, ROR and GRID. One potential solution is to add the metadata without visually displaying it, or creating more types of icons, similar to the ISNI icon to allow links to the icons to be displayed in a visually pleasing manner and does not clutter the screen.

Next Steps

As new content is added to it, there will need to be opportunities to capture this enhanced metadata. The UI of the platform will also need to be improved both to make the related identifiers actionable links and to make the screen more navigable to end users.

The main contributor to data.bl.uk is British Library Labs who develop experimental content based on the British Library's collections. To date they have managed this content and they will continue to do so going forward. A curation workflow will be developed as part of this work to ensure that this metadata is captured for new deposits and that this metadata is sent to DataCite. The possibility of exploring more complex provenance narratives, especially with datasets which are a subset of larger datasets and have undergone a curation workflow, will be explored now that the platform has gone live.

In addition DataCite are working on version 5 of the metadata schema, particularly in relation to incorporating the FAIR principles and aligning contributor roles with other controlled vocabularies such as Data Documentation Initiative (DDI) and CASRAI Contributor Roles Taxonomy (CRediT).⁵ The DataCite Metadata Working Group have also attempted to address the question of organizational affiliations in the latest version 4.3 of their metadata schema.

Ted Habermann, a member of the DataCite Metadata Working Group has carried out some research on affiliations and the benefits of identifiers. Ted notes that the ambiguity of using name strings for affiliations is a principle motivation for identifiers. DataCite includes organizational identifiers for creators and contributors in the latest version 4.3 release of their metadata schema, allowing improved integration of organizational identifiers in DataCite metadata records:

- Addition of new subproperties for Affiliation in the Creator and Contributor properties:
 - affiliationIdentifier
 - affiliationIdentifierScheme
 - schemeURI
- Addition of a new subproperty "schemeURI" for funderIdentifier of the FundingReference property
- Addition of "ROR" to the controlled list values of funderIdentifierType of the FundingReference property.⁶

In research carried out by Ted on textual affiliations he found a case where a single organization was represented five different ways in one metadata collection.

"This problem is pervasive in systems built on free-text entry of any information into any data collection. It is one of the central problems that the introduction of identifiers of any kind is trying to address."⁷

He also found that affiliations written as unaccompanied acronyms are difficult to resolve unambiguously. Again, the use of organizational identifiers can address these issues.

⁵ <https://ddialliance.org/explore-documentation> and <https://casrai.org/credit/>

⁶ <https://www.tedhabermann.com/blog/2019/8/15/metadata-archeology-hunting-affiliations-and-rors-in-datacite-metadata>

⁷ <https://www.tedhabermann.com/blog/2019/11/10/how-many-rors-do-we-need>

Conclusion

This piece of work has been illuminating in illustrating the challenges that are presented in trying to assign and relate PIDs to one another. The provenance use case is also of interest as it is applied to datasets and it will be of interest to see if the suggested interest in this results in actual use by researchers.

Acknowledgements

The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.

References

- Lavasa, A., Dallmeier-Tiessen, S., van de Sandt, S., Dohna, T., Koop-Jakobsen, K., Schindler, U., Ferguson, C., McEntyre, J., Madden, F., Lambert, S., Bunakov, V., Baars, C. (2019). D4.2 Using the PID Graph: Provenance in disciplinary systems. Zenodo. Available online <https://dx.doi.org/10.5281/zenodo.3249833>
- DataCite Metadata Working Group. (2017). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. Available online <https://dx.doi.org/10.5438/0014>