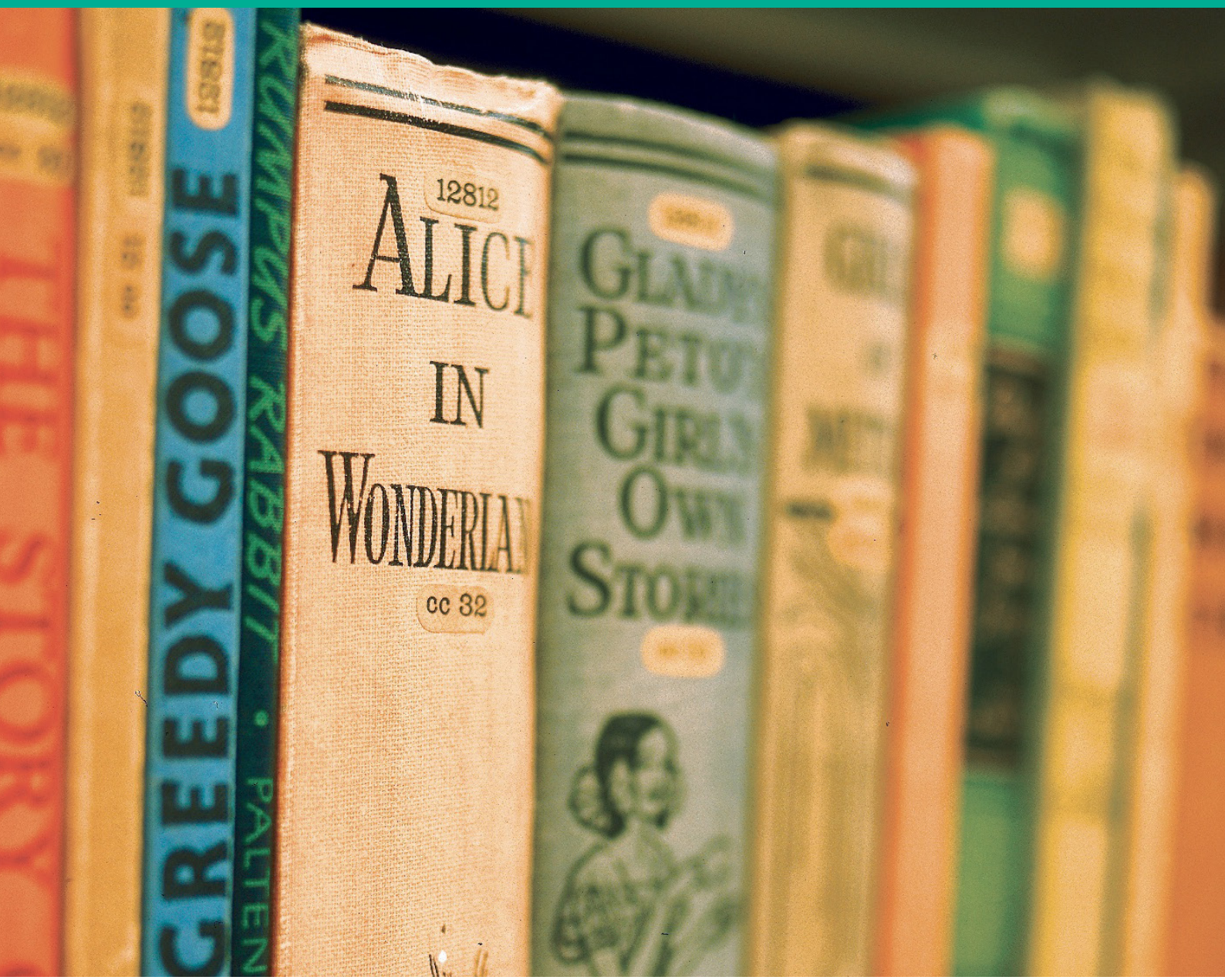
Persistent Identifiers
at the British Library

Frances Madden
orcid.org/0000-0002-5432-6116
Rachael Kotarski
orcid.org/0000-0001-6843-7960



An output of the Persistent Identifiers as IRO Infrastructure project

Arts and Humanities Research Council
TOWARDS A NATIONAL COLLECTION
BRITISH LIBRARY
SCIENCE MUSEUM GROUP
NATURAL HISTORY MUSEUM
V&A
THE NATIONAL GALLERY
Royal Botanic Garden Edinburgh
University of Glasgow

Contents

Executive Summary

This case study provides an overview of persistent identifiers (PIDs) at the British Library, including both those in use and those planned for the future. We hope that this case study will help other heritage organisations to see what the path to use PIDs looks like, and understand what decisions need to be made along that path. Where possible we have tried to demonstrate ways in which organisations can engage and adopt PIDs in their processes and highlight considerations and challenges that may be encountered. In particular we've included the lessons we have learned in facing those challenges, so that our peer organisations can avoid some of those pitfalls.

For the purposes of this case study, a PID is an identifier that is globally unique, actionable (it can be resolved to a user to a resource or information about a resource), and where it is managed to remain unique and actionable for the long term. They may be used to identify the Library's content or elements of its metadata. The systems that create, maintain and hold these identifiers are referred to as PID infrastructure. Within the British Library the main PIDs in use are Archival Resource Keys (ARKs)[1], Digital Object Identifiers (DOIs)[2], International Standard Name Identifiers (ISNIs)[3] and some others including ORCID IDs[4]. Other types of identifiers such as International Standard Book Number (ISBN)[5], International Standard Serial Number (ISSN)[6], shelfmarks and accession numbers are not included in this report, as those identifiers do not meet the definition.

Identifiers for content: Archival Resource Keys

ARKs can be and are used effectively as identifiers for both internal and external users' purposes including managing access to content and citation of material. At the British Library, they are the primary PID and their intended use is to address an internal use case of identifying the digital objects in the Library's collections for internal management. ARKs are free to use but require a resolver service. The resolver can either be built and maintained by the institution, or the N2T resolver[7], maintained by the California Digital Library, can be used instead. ARKs can be used to identify both physical and digital objects, but are only used for digital objects at the Library, due to their initial implementation to address the need of identifying the increased volume of digital materials within the Library's collections. ARKs were selected as the British Library's primary PID because they can be assigned to metadata records as well as digital objects. Since 2012, they have been used to manage and facilitate access to the majority of the Library's digital resources.

The British Library's ARKs are created using a simple custom-designed piece of software, known as the PID infrastructure (PII), and integrate with the Library's strategic catalogue systems via a custom built, metadata extension repository (MER). The decision to design an additional system that sits outside of, but integrates with existing systems made the implementation much easier within the context of the wider organisation and its simplicity has resulted in it requiring little maintenance since it was first implemented in 2012. However, the integration of ARKs across the Library's three main catalogue systems is not perfect and they have not yet been fully implemented as dereferenced and resolvable identifiers as

[1] https://n2t.net/e/ark_ids.html

[2] https://www.doi.org/faq.html

[3] https://isni.org/page/what-is-isni/

[4] https://orcid.org/about

[5] https://www.isbn-international.org/

[6] https://www.issn.org

[7] https://n2t.net/

the internal use case was the priority during implementation. This has limited their use for citation and tracking purposes. While this also means that ARKs do not meet the definition described above, they are included in this case study as they are the Library's primary PID and the intention is to enable external resolvability at a later date.

As systems are upgraded, the issue of maintaining the infrastructure around ARKs becomes more relevant. For example, as the Library's Digital Library System is replaced, the new system has its own internal identifiers, which could theoretically be used for citation but do not have any guarantee of persistence. If the Library were to guarantee the persistence of these identifiers, it would create an additional layer of maintenance. The Library is actively trying to avoid this and focusing on maintaining the use of ARKs as part of the infrastructure and as they are separate technologies, this is easy to do.

For other heritage organisations, ARKs can be implemented in a relatively straightforward manner. They are free to start using and the technical resource required to set up a resolver service can be lightweight. The method used to create the identifier themselves can be similar to the Library's, a UUID or an entirely different method could be used. While opaque identifiers are recommended, they are not required by the standard. Rather than addressing an internal use case, any organisation implementing identifiers should try to ensure that it meets both internal and external users' current and anticipated future needs to maximise the impact of the implementation.

Identifiers for content: Digital Object Identifiers

DOIs are used in several areas of the Library to identify content both purchased and created by the Library. DOIs have identified and provided stable links to research articles for nearly 20 years, and for research datasets and other non-traditional research outputs for 10 years. They are included as fields for published articles within the Library's catalogue where they can be searched. The Library also uses DOIs within its Shared Research Repository system[8], where a DOI look up provides a method to fetch and auto-populate metadata from other systems. There is also functionality to create DOIs for content published by the Library.

Organisations that assign DOIs generally have a membership relationship with either Crossref[9] or DataCite[10]. While the cost of these memberships can vary, depending on the type of membership, they can be prohibitive for smaller organisations. The British Library leads the UK DataCite consortium, which caps the overall cost of using DataCite for organisations within the consortium. DOIs are used for cultural heritage objects, however, their metadata schema is designed for published research outputs or datasets which may make them unsuitable for some heritage uses. DOIs have mandatory minimum metadata, which leads to higher metadata quality but has a resource requirement and can be challenging when describing diverse resources[11].

DOIs for publications that are assigned by publishers can be accommodated easily in repository systems to make managing an institution's outputs or bibliographic resources more manageable for the long term. For those wishing to assign DOIs to their collection items, particularly where the metadata schema is appropriate for the resource, this can be done

[8] https://iro.bl.uk/

[9] https://www.crossref.org/

[10] https://datacite.org

[11] For example, the DataCite metadata schema https://schema.datacite.org/

through becoming a DataCite member, either through the UK consortium or joining DataCite directly. This is particularly appropriate for institutions that already have detailed metadata associated with their collections.

Identifiers in metadata: International Standard Name Identifier

Unlike ARKs and DOIs, ISNIs[12] are not assigned to an institution's holdings but can be used as part of authority control to consistently identify people and organisations associated with a collection or item. The British Library is a founding member of ISNI and is also involved in collaborative cataloguing with other national libraries and contributes to the Program for Cooperative Cataloging's Name Authority Control file (LC-NACO)[13]. To support further adoption of ISNI across the sector, the Library is working to integrate ISNIs in systems used by other organisations and therefore making them more accessible to other organisations.

ISNI is based on VIAF[14] but has a large number of contributors from across the library, archive and music sectors, which means ISNI can provide unified authority control. In 2020, ISNIs are due to be incorporated into the LC-NACO file. Once this work is completed ISNIs will be included in the authority records of all libraries that utilise LC-NACO. The Library is also working to include ISNI identifiers in the linked open data form of the British National Bibliography[15], a list containing all books and journal titles published in the UK and Ireland since 1950. In addition, the British Library is working with publishers to introduce ISNIs into their workflows so that they can be easily incorporated into the Library's catalogue and other organisations' systems. ISNIs are also supported in the British Library's repository for creators and contributors of resources.

ISNI is a membership organisation and there is a cost associated with membership. However, these costs can potentially be offset by savings in authority control resources. For other organisations, a similar approach can be taken of attempting to match existing authority files with ISNI and to begin adding them into authority records as and when possible. This should improve the quality and detail of authority resources. For organisations wishing to create a small number of ISNIs, they can utilise a forthcoming service from the British Library that will create an ISNI for a nominal fee. ISNI membership is available to organisations who want full API access and assignment options.

Other identifiers at the British Library

The Library also encourages the use of identifiers such as ORCID IDs which are designed for use and owned by researchers. ORCID IDs can be included for creators and contributors in the British Library's Shared Research Repository, are widely used within the research sphere and many researchers have and use them as part of their publishing workflows.

[12] https://isni.org/

[13] https://loc.gov/aba/pcc/naco/

[14] https://viaf.org/

[15] https://bnb.data.bl.uk/

Lessons Learned

The British Library was a relatively early adopter of PIDs, so there are aspects of the implementation which could have been done differently or where the decisions made would not be the ones made today.

- The British Library implemented ARKs to address an internal use case and so the external aspect was not implemented due to resource constraints. A wider evaluation of requirements at the outset would have allowed the Library to set out a policy and long-term strategy for use of PIDs early on. This would have enabled the Library to move forward quicker and reap the benefits of persistent links to content sooner.
- Separating PIDs from core systems means that it is much more straightforward to maintain the PID infrastructure when other systems are being replaced. However, this separation does mean careful surveillance is required during upgrades to ensure valuable integrations continue to work.
- The British Library adopted two types of identifier for content, to address different use cases and classes of material. DOIs were considered instead of ARKs but were discounted due to the central control the registration agencies have for DOIs.
- Working cooperatively with other organisations can result in wider benefits and adoption of identifiers. This comes with the burden of the additional time required to manage the increased number of dependencies, as demonstrated in the work to include ISNIs in the LC-NACO file.

Conclusion

PIDs are in wide, but not universal, use throughout the British Library. Their use is somewhat limited to external users due to how they were implemented. These limitations exist, in part due to the reactive nature of the implementations and the difficulties with trying to implement changes across such a large, diverse collection, and the multitude of systems that manage it. The incomplete coverage of PIDs restricts the universal benefits of their use, such as standardised metadata, improved discovery and connecting resources. In the future, the Library plans to adopt a more strategic, policy driven approach to its implementations of PIDs as systems are upgraded in the future and hopes to encourage further citation and usage of PIDs as an access point to their resources. Through gradually integrating PIDs into authority control, and into heterogeneous, rapidly growing collections there is an opportunity to reduce the resources required to understand and navigate collections more easily.

Introduction

Purpose of this case study

This case study aims to provide an overview of persistent identifier (PID) use within the British Library as part of the Towards a National Collection programme to unify collections. It covers the various types of identifiers, both for holdings and content and authority control, and describes what has been implemented or is planned within the Library. It also describes some of the issues encountered and benefits realised through these implementations and provides some indication of the resource required to deliver these services.

Structure of the case study

The case study is structured around the different types of identifiers rather than by collection area. Following a brief introduction to the British Library and PIDs, the case study describes the implementations of various identifiers that are used to manage and access the Library's content and for authority control. There is also a section outlining the gaps and requirements for PIDs that still exist within the Library.

Persistent identifiers

Within the context of this case study a PID is an identifier that is globally unique, actionable (it can be resolved to a resource or information about a resource), and where it is managed to remain unique and actionable for the long term. These features are what bring the most benefits from the use of PIDs in terms of interoperability and feature support. The systems that create, maintain and hold these identifiers are referred to as PID infrastructure.

About the British Library

The British Library is the National Library of the United Kingdom. Established in 1973, it holds over 170 million items including manuscripts, journals, newspapers, sound recordings, videos, stamps, maps, prints and drawings. It is one of six legal deposit libraries in the UK and therefore receives copies of all books produced in the UK and Ireland.

The Library uses a range of identifiers across its collections, but only those which are classified as persistent according to the definition above, or could easily meet this definition are included here, see Figure 1. However, it is worth acknowledging that from the perspective of many users of Library systems, both internal and external, often the 'identifier' used to retrieve and cite information is the shelfmark of the item. The Library would like to change in future and it is one of the reasons for the Library's interest in this project.

Identifiers at the British Library

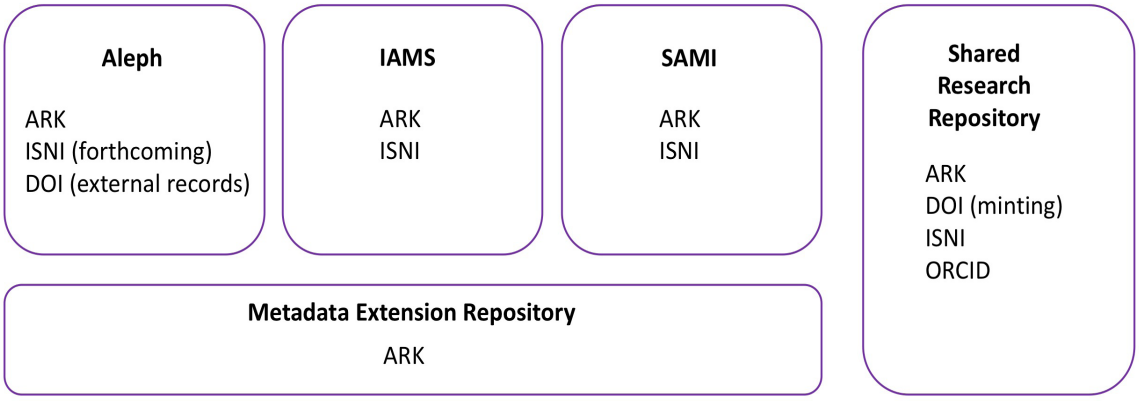


Figure 1: A diagram illustrating the PIDs used at the British Library and the systems in which they are supported. Aleph is the Library's Integrated Library System. IAMS is the Integrated Archives and Manuscripts System. SAMI is the Sound and Moving Image catalogue.

Identifiers for content: Archival Resource Key

Definition and description

Archival Resource Keys (ARKs)[1] were developed in 2001 at California Digital Library. They are designed to provide a resolvable persistent link to any kind of object or metadata about it, both physical and digital.

As of September 2019, it was estimated that over 600 organisations had assigned over 3.2 billion ARKs.[2] The types of records they were assigned to include but are not limited to:

- bibliographic records (15 million, BnF main catalogue[3])
- museum specimens (11 million, Smithsonian[4])
- digitised documents and objects (5 million, BnF Gallica[5])
- historical authors and scholars (4 million, SNAC[6])
- finding aids and special collections (4 million, Merritt[7])

ARKs are a string of characters including the label "ark:", see Figure 2. They can be expressed as a URL with the name of the service supporting that ARK preceding it and, in fact, this is the preferred format for citation, see Figure 3. The service name can change but the ARK itself cannot. The ARK contains a name assigning authority number that indicates which organisation created the ARK.

[1] https://n2t.net/e/ark_ids.html

[2] https://wiki.lyrasis.org/display/ARKs/ARK+Identifiers+FAQ#ARKIdentifiersFAQ-ARKedThingsWhatkindsofthingsareARKsassignedto?

[3] https://catalogue.bnf.fr/

[4] https://library.si.edu/

[5] https://gallica.bnf.fr/GallicaEnChiffres

[6] https://snaccooperative.org/

[7] https://merritt.cdlib.org/. Self reported figures, correct at September 2019.

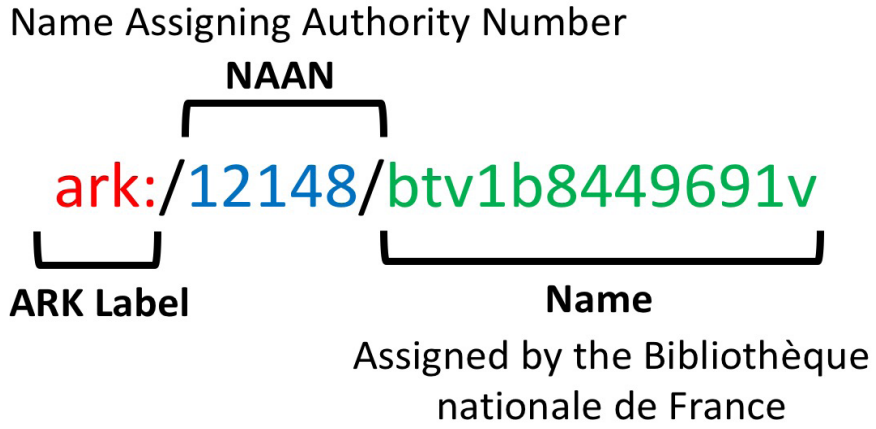


Figure 2: Format of an ARK. The ARK Name can be opaque or non-opaque.

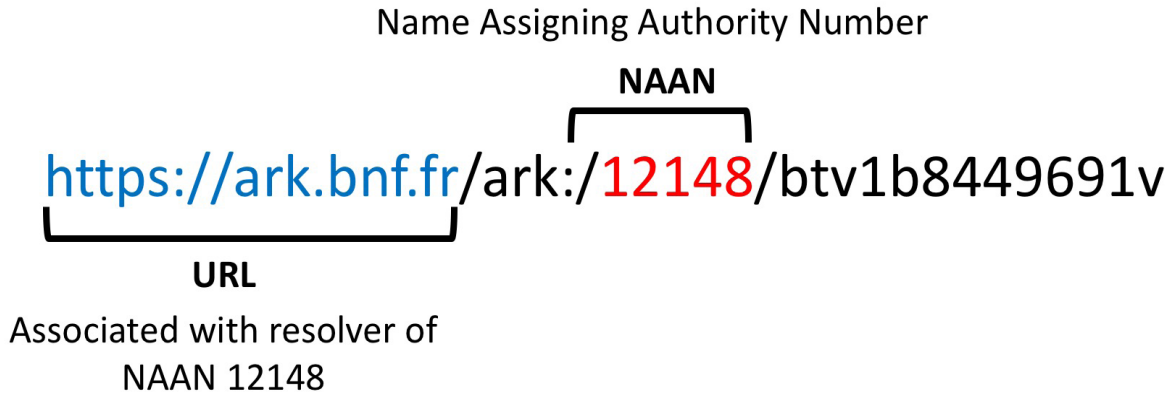


Figure 3: An ARK as a URL

ARKs are free to use and implement. If done internally, this would require internal technical support to manage their assignment and resolution, but the centralised N2T resolver, provided by California Digital Library can be used instead.[8] In order to start using ARKs, one needs to register for a Name Assigning Authority Number (NAAN) that can be obtained free of charge. They are a federated standard and the assigning institution has a high level of control over the ARKs they assign. They have no metadata requirements associated with them and can be used for both internal administrative and external discovery services. ARKs differ from other identifiers in that they can be deleted and reused if they refer to a resource which no longer exists and has never been made public.

ARKs at the British Library

ARKs are the primary PID in use for content at the British Library. ARKs are used to identify a wide variety of objects including digitised materials, e-books, metadata records, sound and audiovisual materials. They were conceived as providing identifiers for digital objects only and their implementation was designed to address the internal use case of managing a large and fast-growing digital collection. They are used to manage and facilitate access to the majority of the Library's digital resources.

The structure of a British Library ARK is outlined in Figure 4. The Library's systems are configured to create ARKs composed of three different components as described in Figure 5, and they are always entirely opaque, that is the ID itself contains no semantic information.

[8] https://n2t.net/

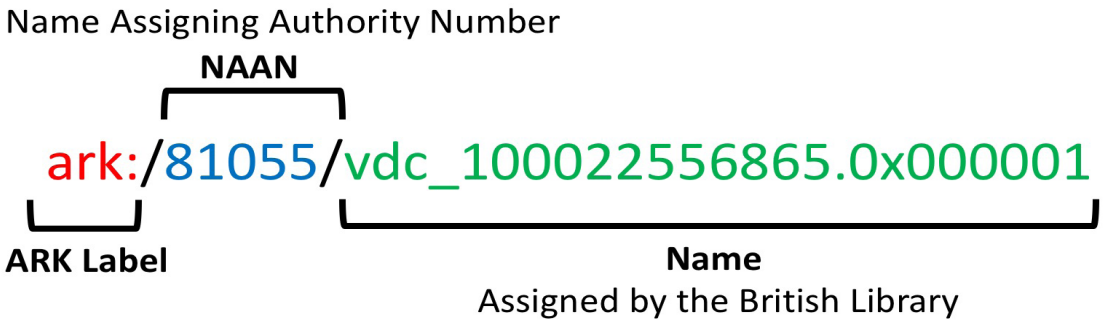


Figure 4: Format of a British Library ARK

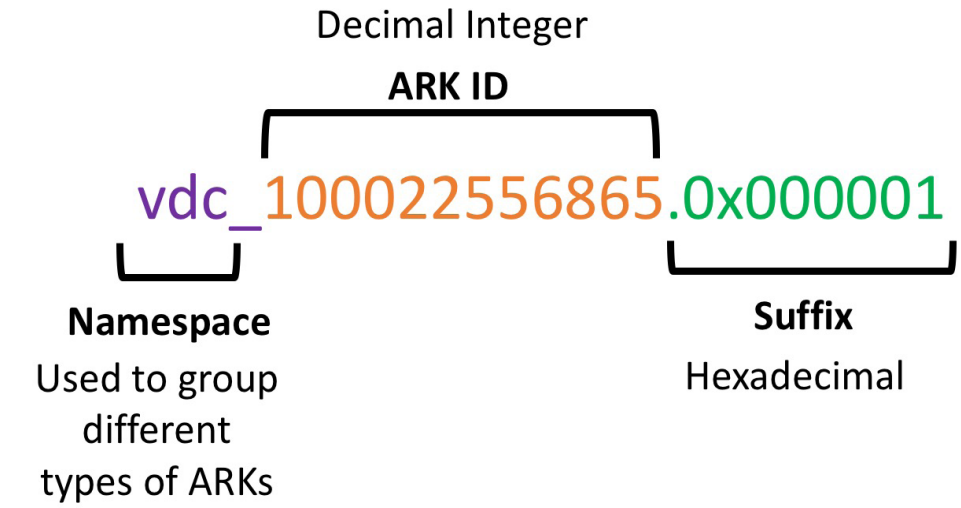


Figure 5: British Library ARK Structure

The British Library's core cataloguing systems cannot manage all the metadata the Library requires, including PIDs, so a metadata extension repository was built to accommodate this additional metadata. ARKs are stored in the metadata extension repository and are added to the Digital Library System (DLS) and metadata database as part of a METS[9] file, which is used for access and preservation of digital objects.

The Library has internally defined different types of ARKs that are used depending on how the digital object it identifies sits within the structure of the system. As ARKs are opaque, it is not possible to identify from the ARK itself what type it is, but these types are used in workflow design and to help understand what ARKs are associated with a given record or resource.

- MD-ARK – Metadata ARK. Identifies a metadata record
- L-ARK – Logical ARK. Identifies a logical entity, e.g. a book, an article. Comprised of groups of D-ARKs, P-ARKs or SL-ARKs
- D-ARK – Digital ARK. Identifies individual digital objects
- P-ARK – Physical ARK. Identifies parts that make up a logical entity e.g. pages of a book, areas on a newspaper page, one side of a vinyl record
- SL-ARK – Sub logical ARK. Identifies an L-ARK that is the child of one or more L-ARKs.

[9] Metadata Encoding and Transmission Standard - https://www.loc.gov/standards/mets/

ARKs are created for the metadata records of digitised materials in the published catalogue Aleph,[10] for all metadata records in the archives and manuscripts catalogue IAMS and the sound and moving image catalogue, SAMI. All digital objects, regardless of which catalogue they are described, have ARKs when ingested into the DLS. For end users of the Library viewing digital content accessible via the Universal Viewer, the L-ARK is the one PID that is exposed to users in the form of a URL for a resource and is included in the suggested citation URL, see Figure 5. For users accessing other types of digital content, the L-ARK is either not visible or briefly visible and is not a practical citation link, again due to the internal use case for which this process was designed.

Why the British Library uses ARKs

ARKs have been in use at the Library since 2012. In 2009/10, a detailed analysis was undertaken of the various types of identifiers in use and available at the Library at that time and no one identifier entirely met the British Library's requirements. All of the identifier schemas which the Library had in use were tied to special content types or implied a certain underlying content model which is not universally adaptable. Two main 'universal' PID types were considered, ARKs and Handles[11]. ARKs were selected once it was determined it was possible to assign ARKs to an object's metadata. It had been identified that each metadata entity needed to have a PID, which was not described in any of the recommended implementations of ARKs to that point in 2011 and therefore the Library needed to adapt their implementation to accommodate this requirement of assigning to metadata entities.

Description of implementation

The PID infrastructure (PII) creates ARKs at the British Library and is used for matching records which are loaded into Aleph and other databases. It consists of a custom built SQL database and a javascript application that queries the database to generate the ARKs which are assigned to entities. An algorithm is used to generate opaque identifiers and to ensure they do not repeat, the integer value increments upward for each request and within each request for each item the hexadecimal value increments upward. The ARKs are included as part of a METS file sent to the metadata extension repository (MER), a custom built repository for metadata that cannot be accommodated elsewhere, and ingested into the DLS as part of a submission information package (SIP).

According to the ARK standard, ARKs can be deleted if they have not been published anywhere[12]. However, at the British Library a retirement process is used where the status of the ARK in the PII changes to indicate it is no longer active but it is not deleted from the database.

Both the PII and MER are systems developed internally at the Library. The PII was developed within six months by one developer. The system is very simple and has been running since 2012 with no need for any maintenance or development, and could be made available to other institutions wishing to assign ARKs. MER was developed around the same time and took a developer approximately one year to develop. Again, this system is very robust and has not needed any maintenance since 2012. These systems were relatively straightforward to develop and had a much lower overhead than overhauling the Library's strategic systems, while still developing the Library's capability to create and manage PIDs.

Benefits and challenges in using ARKs

Due to the simple and lightweight approach taken to assigning PIDs to the Library's digital

[10] https://www.exlibrisgroup.com/products/aleph-integrated-library-system/

[11] https://www.handle.net/

[12] https://wiki.lyrasis.org/display/ARKs/ARK+Identifiers+FAQ#ARKIdentifiersFAQ-MoredifferencesbetweenARKs,DOIs,Handles,PURLs,andURNs

objects, it would also be simple to change to using another type of identifier. The PII is a discrete piece of architecture and could be changed for something else or adapted. The most important aspect of whatever technology used is that it generates truly (globally) unique identifiers.

ARKs were implemented at the British Library and integrated with several different systems. This was a pragmatic approach recognising that PIDs were needed quickly to accommodate the growing digital collections but the existing systems could not manage them.

As they integrate with several systems, it is unsurprising that some of these integrations are imperfect. The Library's core cataloguing systems are long standing strategic systems and therefore can be challenging to integrate. For example, the archive and manuscripts system, IAMS, is configured to store a metadata ARK for every record as soon as it is created. However due to the restrictions of that system it is not possible to change the type of the record, e.g. from item to file or file to series, and it has to be deleted and a new one created. This can be problematic due to human error, but also during projects where collections are catalogued in greater detail, and the hierarchy of collections might change. The issue is the decision to retire ARKs where needed, which is an extra process requiring additional technical resources.

With the British Library's focus on near-term internal requirements, certain aspects of the standard[13] were not implemented in the resolver due to resource constraints.[14] In addition, the Library's original implementation design included functionality to specify delivery services or rendering agents from the catalogue record, i.e. determining which type of service, an e-book reader or an audiovisual player should render the object, but only one 'https://access.bl.uk' service was implemented. This has consequences as systems are upgraded to support more diverse materials, such as audiovisual content, which may require different services.

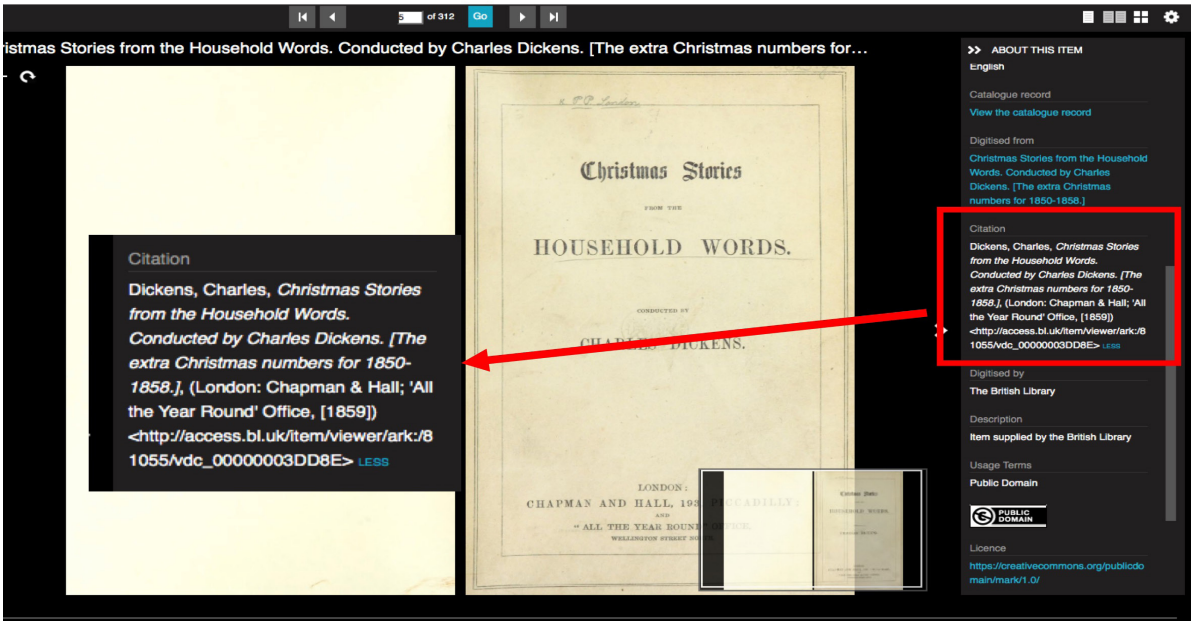


Figure 6: A suggested citation for a digitised resource. The ARK is included in the citation URL.

[13] Such as 'inflections' e.g., appending a '?' to an ARK returning a brief human and machine readable metadata statement for that ARK

[14] ARKs in Action - Inflections https://n2t.net/e/ark_ids.html

Another element of the ARK standard that was never fully implemented is the ability for users external to the Library's network to resolve ARKs. Again this is due to the fact it was implemented to address an internal use case and resources were constrained. For digitised materials, the ARKs are used as a component for the URL but are not resolvable in and of themselves via the standard URL syntax. While this means that ARKs do not fully meet the definition described at the beginning of this document, they are included as they are the main PID used by the Library and could easily be configured to meet this definition. This is unfortunate as the format used makes the URL very long, discouraging citation of the digitised resource in full and the ARKs cannot be resolved through the universal N2T resolver, maintained by California Digital Library. The drawbacks of this is that it is difficult to track usage of the Library's digitised collections, see Figure 6, and potentially inhibit access to a resource if anyone tries to access the ARK without its resolver URL. This may be addressed in future as part of a programme to upgrade the Library's core systems.

Another challenge of continuing to implement ARKs or any PID for objects is the need to guarantee persistence through changes in the underlying infrastructure. One major benefit of PIDs is that they offer a resolver service which dereferences the identifier from its URL so URLs only have to be updated within the resolver, in this case within the Library's access stack. The Library is in the process of replacing its underlying digital preservation solution, the Digital Library System. This new system has its own internal identifiers, however the level at which these identifiers are assigned is at a slightly different level of granularity to which the British Library has assigned ARKs and they do not guarantee persistence beyond the lifetime of that particular system's use. The Library required that this new system be able to support ARKs, which it does, however these internal identifiers will be generated for records within the system and theoretically could be surfaced to end users as a citation mechanism. If they were, the Library would need to guarantee their persistence, adding an additional maintenance requirement. It is intended ARKs will be maintained and included within the implementation of this new solution because they may have been cited by readers in the past and the Library deems it necessary to maintain them for the future, independent of other infrastructural changes. As systems evolve, institutional memory and clear policies are required to ensure the continued maintenance and persistence of identifiers. It is easier and preferable to focus on one identifier for a given item for the purposes of resolution and citation to reduce the maintenance burden on the Library in the future.

For other heritage organisations, ARKs can be implemented in a relatively straightforward manner. They are free to use and the technical resource required to set up a resolver service can be reasonably lightweight. The method used to create the identifier themselves can be similar to the Library's, a UUID or an entirely different method could be used. Rather than addressing an internal use case, any organisation implementing identifiers should try to ensure that it meets both internal and external users' now and anticipated future needs to maximise the impact of the implementation.

Identifiers for content: Digital Object Identifier

Definition and description

Digital Object Identifiers (DOIs) are an identifier used to identify a wide variety of digital objects. Based on Handle technology, it was adopted as ISO 26324 in 2012 and it is governed by the International DOI Foundation[15]. DOIs extend the functionality of Handles in that they require metadata and governance that supports further persistence of the identifiers. Registration Agencies manage DOIs, and users of these registration agencies are assigned a prefix string, to which they add a unique suffix to form a new DOI, see Figure 7. The prefix ensures that as long as a suffix is unique to the prefix, the whole identifier will be globally unique. DOIs can be expressed as URLs, see Figure 8.

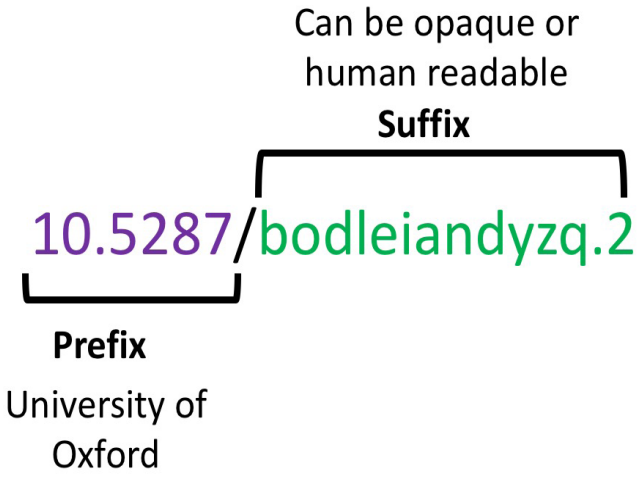


Figure 7: The structure of a DOI

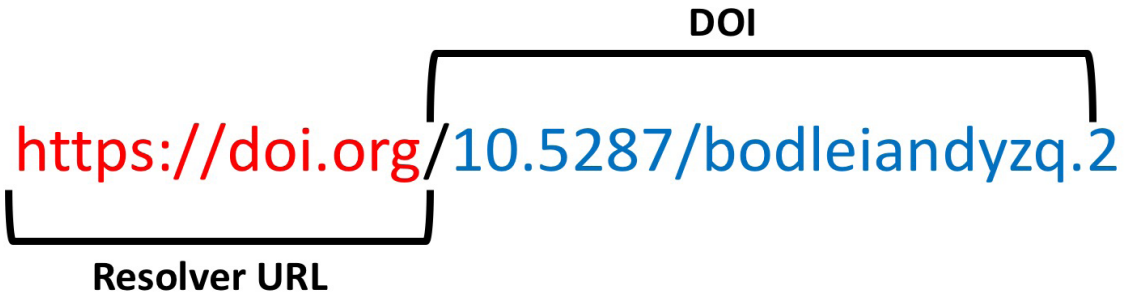


Figure 8: A DOI expressed as a URL

Crossref[16] and DataCite[17] are the two DOI registration agencies that assign DOIs to research produced in the English speaking world. Crossref has been in existence since 2000, and is primarily concerned with published outputs such as journal articles, book chapters and books. DataCite, founded in 2009, was conceived initially to provide DOIs for research datasets but this has expanded in recent years to include all types of non-traditional research outputs, e.g. theses, grey literature and software.

DOIs have a defined minimal set of metadata fields whose inclusion is mandatory as part of the assignment process. For DataCite DOIs, this is limited to six fields.

[15] https://www.doi.org/

[16] https://www.crossref.org/

[17] https://datacite.org/

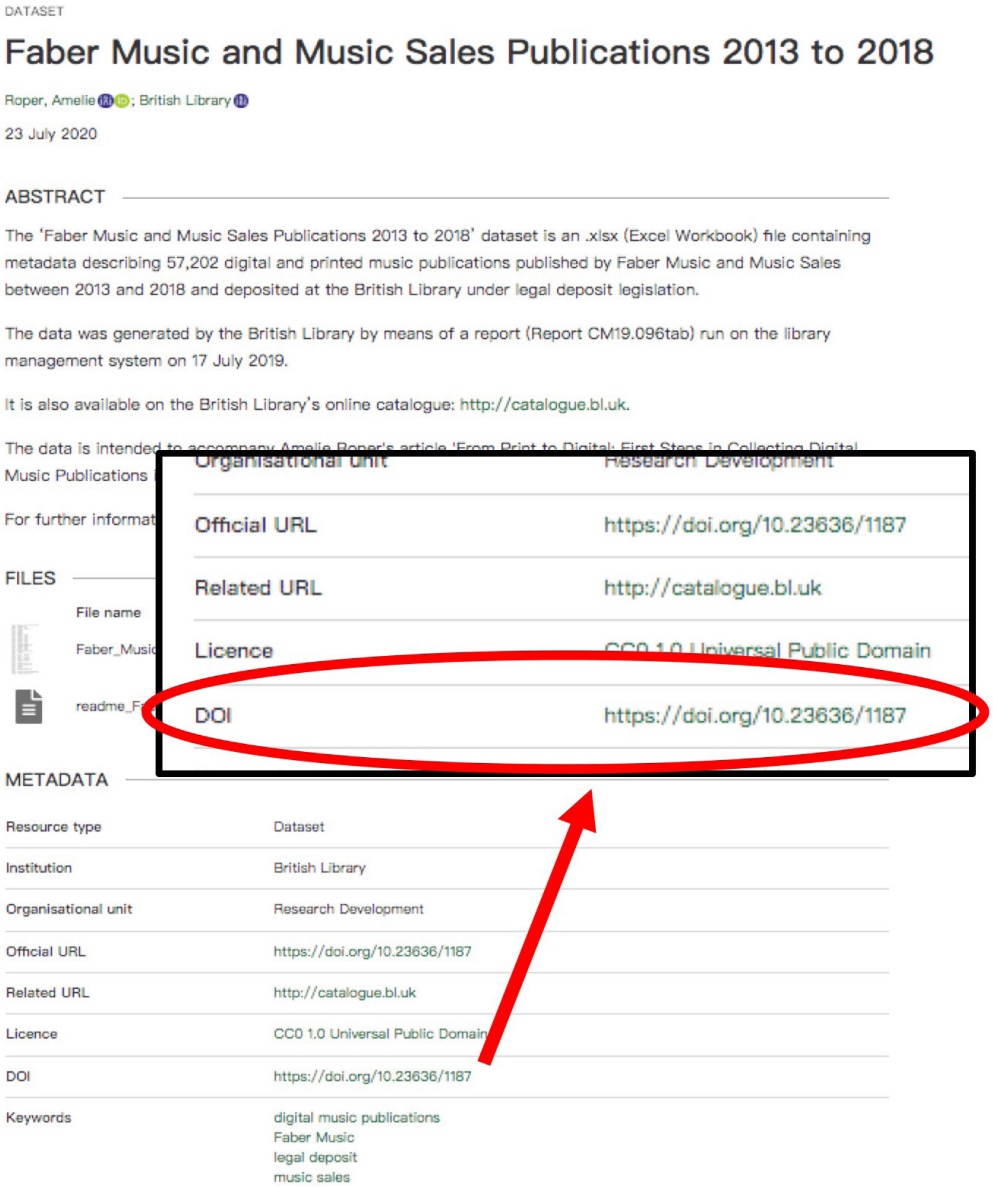


Figure 9: A record from the British Library's Shared Research Repository highlighting its DOI that was created through the repository's administration interface.

DOIs at the British Library

As DOIs are widely assigned for published material, they are included in the Library's catalogue for published journal articles which are purchased from publishers or acquired via Legal Deposit. DOIs are used within the British Library's Shared Research Repository[18] to identify both the Library's own research outputs, and datasets derived from its collections, mostly created by the British Library Labs team. DOIs are used in the repository as the content has potential to be reused as research material in its own right and having a DOI increases the discovery potential of the output. The major benefit to using a repository for creating and managing DOIs is the central metadata store. The repository software allows you to fetch metadata from Crossref's and DataCite's central metadata stores and populate the metadata fields in a repository record. It also allows you to create DOIs for records that have been added to the repository and for which the repository is their main publication outlet, see Figure 9.

[18] https://iro.bl.uk/

Stemming from an interest in research reproducibility and an attempt to support the collection of the whole research record of the UK, the British Library was a founding member of DataCite and has supported it from its inception. As such, the British Library is the DataCite UK consortium lead, where it administers the DOI prefixes assigned to UK research institutions in the consortium and promotes the use of DataCite DOIs in the UK.

Benefits and challenges of DOIs
DOIs are a well-established standard for published research outputs and research datasets. More recently, they have become the accepted means to provide persistent citable identifiers to other non-traditional research outputs including theses, grey literature and reports.

Depending on the nature of the resource that needs an identifier, DOIs may be more or less suitable. In particular, each Registration Agency has metadata requirements specialised for particular types of resource. The DataCite metadata schema has evolved over time to support a wider array of research output types beyond research data. Crossref's metadata scheme is designed to describe published outputs and therefore best meets that community's needs. The Entertainment Identity Registry Association (EIDR)[19] is a DOI registration agency but their infrastructure is designed for film and television assets and have a metadata structure designed to meet their community's needs.

DOIs were considered in place of ARKs for assigning to individual collection items at the British Library, but were discounted due to the level of central control from DOI registration agencies in contrast to ARKs' distributed infrastructure where the Library has authority. There were concerns that any of the registration agencies would not agree to the variety of content for which the Library wished to assign identifiers and that the library would still need to manage the infrastructure to create unique suffixes for its DOIs. At the time, DataCite DOIs operated primarily in native XML and did not provide support for RDF[20], which the Library wished to support, as that is the format used in MER.

Generally assigning DOIs depends on the membership fees of a registration agency and the types of membership offered. These costs can vary across institutions and depend on the number of DOIs created. For example, from 2021 members of the UK DataCite consortium will pay approximately £1,000 per year.

Accommodating the DOIs created by publishers for articles and other published items in repository systems will make managing the institution's own research outputs or bibliographic resources more manageable for the long term. For those wishing to assign DOIs to their collection items, particularly where the metadata schema[21] is appropriate for the resource, this can be done through becoming a DataCite member, either through the UK consortium or joining DataCite directly. This is particularly appropriate for institutions which already have detailed metadata associated with their collections. Members need to be able to guarantee to maintain the DOIs so they remain resolvable and accessible for the long term.

[19] https://eidr.org/

[20] https://www.w3.org/RDF/

[21] For example, the DataCite metadata schema https://schema.datacite.org/

Identifiers for metadata: International Standard Name Identifier

Definition and description

The International Standard Name Identifier (ISNI)[22] is an ISO standard identifier (ISO 27729) which can be used to identify contributors to creative works including but not limited to researchers, writers, artists, performers, publishers and aggregators. It aims to assign the public identity of an individual or organisation a persistent unique identifying number to disambiguate names and aid search and discovery. ISNI is a sixteen digit identifier which can be expressed within a resolvable URL, as shown in Figure 10.

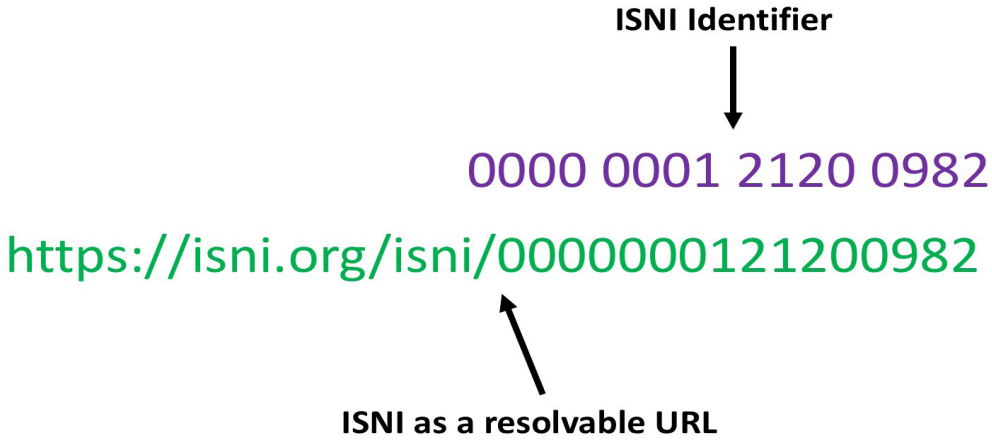


Figure 10: Format of an ISNI

ISNI and authority control

ISNI is different from other types of authority control as it is centrally managed via a database hosted by OCLC. Unlike traditional, locally maintained authority control, ISNI is designed as a bridge identifier providing a central location to which authority records can be linked. In turn, this reduces the authority control resource requirement on an individual library. The original ISNI database was created using data from the Virtual International Authority File (VIAF)[23] so its contents tend to suit library collections. However as ISNI has grown, it has registration agencies from across the music industry, archives and the research sphere resulting in greater diversity in the database and coverage of previously proprietary silos. ISNI uses matching algorithms and assigns confidence levels to match new names to existing ones to ensure the authoritative assignment of identifiers.

ISNI and the British Library

The British Library is a founding member of ISNI, a member of the ISNI-International Agency Board and a registration agency for the standard. The Authority Control team based at the British Library creates ISNIs and matches ISNIs and names, both manually and on a batch basis to ensure quality control, maintaining and enhancing the database. The wider team at the British Library works on behalf of ISNI to drive adoption across sectors, some benefits of which are due to be realised soon. For example, it is engaged in working with the publishing industry to adopt ISNI into publishing workflows, which is envisaged will reduce authority control effort for many organisations in the future.

Existing integrations

ISNI is supported within the British Library's Shared Research Repository where they can be included for both individual and organisational creators and contributors of works, see Figure 11. ISNIs are also included for authors in EThOS[24], the UK's national index of theses,

[22] https://isni.org/

[23] https://viaf.org/

[24] https://ethos.bl.uk

shown in Figure 12. These are created via a batch process by OCLC and are quality assured by the Authority Control team at the British Library. Authors of PhD theses are found to be good candidates for creation by a batch process as a PhD thesis is often an individual's first published output, so they are unlikely to have had an identifier previously assigned.

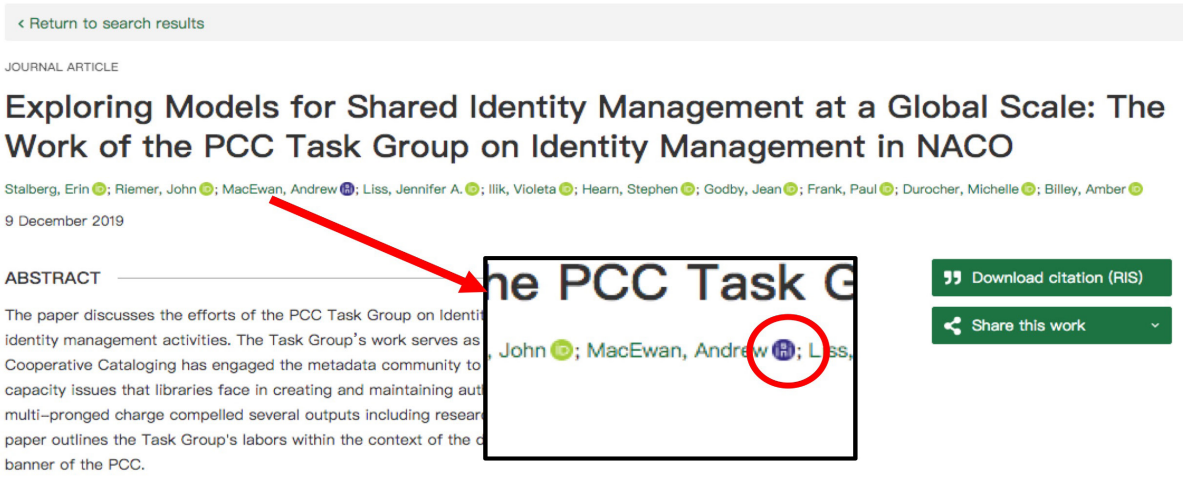


Figure 11: ISNI in the British Library's Shared Research Repository. The blue icon is an actionable link to the ISNI record.

ISNI can also be included within the archives and manuscripts catalogue, IAMS, as an identifier within an authority record. However, ISNIs are added on a limited, best efforts basis for example as records are deduplicated through normal business processes or as an approach by individual digitisation or cataloguing projects. As such they are included in less than one per cent of records.

Figure 12: An ISNI in an EThOS thesis record. The ISNIs are created through a batch process and displayed as resolvable links in the record.

Planned integrations

The British Library has taken a cooperative approach to authority files for many years and is a long standing contributor to the Program for Cooperative Cataloging's (PCC) programs for name authority control (NACO)[25] and subject authority control (SACO)[26]. In 2020, ISNIs will be integrated into the NACO file using a gradual process so the file is continually accessible. Since the Library's main catalogue system uses the NACO file in its records, these will be linked to ISNIs and they will be available to all other users of the NACO file too. ISNI will be further integrated into IAMS, which manages the Library's unpublished collections and SAMI, the sound and moving image system through machine processing and curatorial activity.

[25] https://www.loc.gov/aba/pcc/naco/

[26] https://www.loc.gov/aba/pcc/saco/

The Library is also working to include ISNI in the British National Bibliography, the backfile of bibliographic records for all books and journals published in the UK and Ireland since 1950 which is available in linked open data format. Some automated checking of authors has already taken place, of 4.6 million names; 3.1 million could be matched to ISNIs.

By automating the process as much as possible, it is hoped that this process can be cost neutral, though final costs will not be clear until the workflow is finalised. The work with publishers to encourage them to utilise ISNI will also help to offset this cost in the future.

EThOS is due to be migrated to a new platform in the coming years and at that point, there will be an opportunity to include further identifiers, including ISNIs for PhD supervisors and organisations such as the awarding institution, and any sponsor of the research.

Benefits and challenges of using ISNI
Traditional authority control measures are costly in terms of staff resource and the growth in collections, particularly of Library collections, may make a more automated process attractive to organisations. The Library's integration approach for ISNI is incremental but it maintains continuity of practice within existing workflows, enables more effective exploitation of authority data for discovery and lays the foundations for cross collection searching and audience focused discovery. It also does not require extensive changes to existing data. While ISNI membership does have a cost, it facilitates access to a large database as well as tools to enable a high quality of authority control data for a collection and introduces the potential to offset against existing costs.

Some cultural heritage organisations experimenting with assigning ISNI to their collections include the Bodleian Medieval Manuscripts, who have matched their authority files with ISNI.[27] For other organisations, a similar approach can be taken of attempting to match existing authority files with ISNI and to begin adding them into authority records as and when possible. This should improve the quality and detail of authority resources. For organisations wishing to create only a small number of ISNIs, they can utilise a new service from the British Library which will create an ISNI for a nominal fee. ISNI membership is available to organisations who want full API access and assignment options.

Other PIDs at the British Library

ORCID IDs[28] are persistent unique identifiers designed for researchers, and in contrast to ISNI, they are owned and controlled by the researchers themselves. They are designed to help disambiguate between researchers and can be used in publication management systems and resource discovery. They are sixteen-digit unique numbers that are resolvable as URLs in the format, https://orcid.org/0000-0001-6480-7047, the identifiers themselves being a subset of ISNI. The British Library's Shared Research Repository includes functionality for including the ORCID ID of a creator or contributor and this appears as a resolvable link in the repository's record, shown in Figure 13. ORCID IDs are also supported within EThOS and are included as actionable links in records where they can be harvested from universities.

[27] For example https://medieval.bodleian.ox.ac.uk/catalog/person_9966678

[28] https://orcid.org/

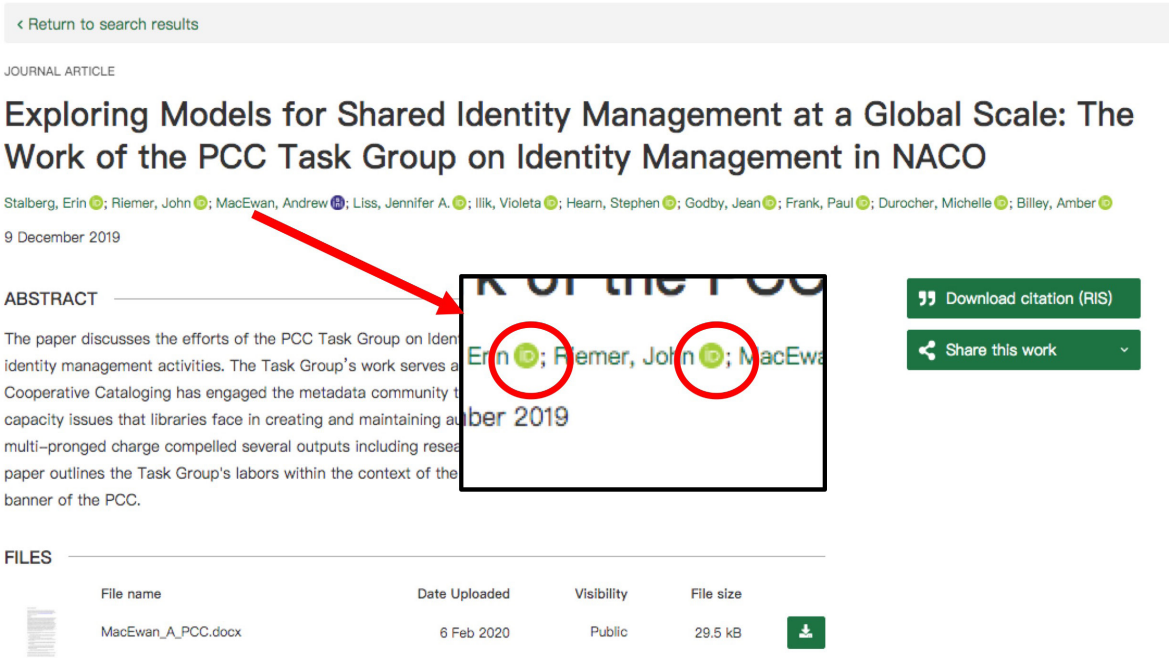


Figure 13: ORCID is supported in the repository where they are displayed as green actionable icons which resolve to the creator's ORCID profile.

As part of the European Commission funded FREYA project[29], support for emerging types of identifiers is being incorporated into the repository including Research Organization Registry (ROR)[30] identifiers for organisational contributors and creators, Crossref Funder Registry[31] identifiers for funders, and GRID[32] and Wikidata[33] identifiers for organisational contributors and creators.

Gaps - where PIDs are still needed

While the Library has implemented PIDs for many parts of its collections, there are still some areas that do not have a persistent resolvable identifier, e.g. print books in the collection do not have a persistent resolvable identifier to their metadata record. In part, this is due to the structure of the Aleph cataloguing system, which assigns system identifiers that are not necessarily persistent, to the title and holding of a record but not at the item level. It is anticipated this issue will be addressed through an ongoing change programme which is tasked with upgrading the Library's systems.

Equally, items within other collections areas such as archives and manuscripts only have an ARK if they have been digitised, as ARKs are assigned to digital materials only. Within the archives and manuscript catalogue, IAMS, the records have metadata ARKs but the items themselves (unless digitised) do not. This means the ARKs are not visible to an end user and that citation using persistent resolvable identifiers is not possible.

News Collection

The curators managing the British Library's newspaper collection have also identified that consistent authority control using identifiers could have a major benefit to their collection management and curatorial processes. The newspaper collection is large and heterogeneous.

[29] https://www.project-freya.eu/en

[30] https://ror.org/

[31] https://www.crossref.org/services/funder-registry/

[32] https://www.grid.ac/

[33] https://www.wikidata.org/wiki/Wikidata:Identifiers

It spans 400 years and includes approximately 60 million issues and 33,000 titles, has large amounts of content digitised and is affected by Legal Deposit Legislation[34]. Various kinds of identifiers have been used over time across parts of the collection but nothing has been used consistently for all types of records. Newspapers often change title, in some cases many times, e.g. the London newspaper, The Evening Standard, has had ten titles since the nineteenth century. These changes are managed in Aleph by creating a new record for the title on each occasion. It is often challenging for curators and researchers to navigate across these and understand how different titles relate to each other.

The changes in the news media industry and the reduced primacy of the print edition mean that managing each issue and their versions has become increasingly challenging and the Library's systems need to be updated to deal with this change. The system was conceived for an analogue world but collecting now incorporates digital news in several forms, including web and broadcast. Due to the current Legal Deposit Legislation, the Library only collects the print edition of most papers, rather than both the online and the print. As newspaper titles move towards becoming brands with multiple manifestations for different audiences and platforms, PIDs could help the Library manage this systematically, especially if the newspapers themselves were to assign the identifiers.

Lessons Learned

The British Library has been using PIDs for nearly a decade. As it was a relatively early adopter of PIDs, it has learned several lessons and there are aspects of the implementations which could be done differently, particularly now that identifier schemes have developed.

When implementing ARKs, addressing an internal use case only was done to meet a pressing institutional need at that time and the external use case was removed from scope. Making ARKs externally resolvable has always struggled to gain priority since then and has not yet been achieved. While it was certainly pragmatic and helpful to implement a solution for PIDs which did not require the Library's core cataloguing systems to be overhauled, the imperfect integrations between the systems does restrict the extent of the use of PIDs, particularly for non-digital objects, and it has also been challenging to secure resource to improve these integrations over time.

As systems are upgraded over time, the fact that the PID infrastructure sits separately from the rest of the core cataloguing systems makes it much easier to maintain the PIDs and focus on integrating with the new systems instead.

The British Library made the decision to adopt two different types of identifier to address the divergent use cases for different types of content. ARKs are assigned to digital objects within the main collection and DOIs are assigned to research outputs either derived from the Library's collections or created by staff and associates of the Library. The identifiers are often assigned at different levels of granularity, e.g. a dataset composed of digitised materials has a DOI but every item within the dataset would have one or more ARKs. In addition, the infrastructure around DOIs has changed and developed a lot since the Library decided to implement ARKs. DataCite's metadata schema has expanded to accommodate a broader range of entities including types of heritage objects. While the ARK and DOI creation functionality is entirely separate at the Library, the identifier creation functionality could be unified where a unique suffix is created and appended either to an ARK or to a DOI.
The Library's work with integrating ISNI has demonstrated the benefits of working cooperatively across organisations. While arguably it may have taken longer to integrate ISNI into the Library's catalogue by doing it via the LC NACO file, it has meant that the benefits will be shared with other organisations and that a more cohesive approach is adopted

[34] https://www.bl.uk/legal-deposit/about-legal-deposit/

by the contributing organisations. A similar benefit could also be seen with UK heritage organisations in their adoption of PIDs if such an approach were taken.

Over time, the Library's approach towards PIDs has become less reactive and they have become a core part of the Library's infrastructure, this requires a strategic approach, which the Library needs to implement, especially as systems are migrated and developed.

Conclusion

While PIDs have been used in some areas of the British Library for some time, there are many areas in which the Library is trying to improve its practices and integrate existing and emerging identifiers into its systems. The Library's existing implementations of PIDs have generally been in response to user and organisational needs, and evolved organically over time, often while the identifiers themselves have also evolved. The size of the Library's collection and the range of systems used to manage it, are challenging for making universal changes. There is a need to streamline and be strategic in future implementations, particularly as the Library's wider change programme develops and plans replacements of the existing core library management systems. The Library will develop a policy on PID use across the organisation to standardise its approach across the organisation.

While the PIDs the Library does use enable long lasting linking and citation of resources, this potential benefit is not realised fully. In the future, the Library hopes to encourage further citation and usage of PIDs as an access point to their resources. Through integrating PIDs into authority control, and into heterogeneous, rapidly growing collections such as the news collection, there is an opportunity to reduce the resources required to understand and navigate collections more easily.

Glossary

- Aleph – Library catalogue system for published material
- ARK – Archival Resource Key
- DOI – Digital Object Identifier
- EThOS - the UK's national theses service based at the British Library
- IAMS – Library catalogue system for unpublished material, e.g. archives, manuscripts, visual arts, maps, philatelic collections etc.
- ISNI – International Standard Name Identifier
- ORCID – Open Researcher and Contributor Identifier
- SAMI – Sound and moving image catalogue
- Shared Research Repository - a repository for cultural and heritage organisations administered by the British Library based on Samvera Hyku.

Appendices

Appendix 1 - British Library Staff Consulted

- Jenny Basford, Repository Services
- John Beaman, Digital Preservation
- Fiona Clancy, Heritage Made Digital
- Jez Cope, Data Services
- Paul Clements, Technology
- Ian Davis, Acquisitions and Cataloguing South
- Sara Gould, Repository Services
- Andrew MacEwan, Collection Metadata
- Peter May, Digital Preservation
- Luke McKernan, News and Moving Image
- Nicolas Moretto, Collection Metadata
- Emma Rogoz, Collection Metadata

Appendix 2 - About Persistent Identifiers as IRO Infrastructure

Museums, heritage collections and sites in the UK house at least 200 million physical and digital objects. Being able to identify these objects supports their discovery, use and curation – you cannot provide persistent or even consistent access to an item if you don't know what it is. Accession numbers are a key component in all collection and library management systems but these only cover selected objects within an individual collection. To fully realise the potential of our national collections, we need identifiers that can bring together collections across institutional boundaries.

Persistent Identifiers (PIDs) provide a long-lasting click-able link to a digital object. They are recognised by UKRI as a tool for enabling data discovery, access and citation. Supporting wider use of PIDs for collection objects, environments, specimens and related items will allow long-term, unambiguous linking of collections that will create a digital National Collection. However, the challenges, utility and wider benefits of PIDs are not as well understood across the heritage sector as they could be.

This project will bring together best practices in the use of PIDs, building on existing work and projects. We will share expertise and provide recommendations on the approach to PIDs for colleagues in institutions across the UK heritage sector. Through a mixture of workshops, desk research and case studies, the project will answer questions such as 'What are the gaps in the existing PID landscape for heritage collections, buildings and environments?' and 'What should a PID infrastructure, strategy and governance framework look like for a unified national collection?'.

This project is a Foundation project within the AHRC funded Towards a National Collection Programme.[35]

Appendix 3 - About Towards a National Collection

Towards a National Collection is a major five-year £18.9 million investment in the UK's world-renowned museums, archives, libraries and galleries. Funding is provided through UK Research and Innovation's Strategic Priorities Fund and delivered by the Arts and Humanities Research Council (AHRC). The programme will take the first steps towards creating a unified virtual 'national collection' by dissolving barriers between different collections – opening UK heritage to the world. By seizing the opportunity presented by new digital technology, it will allow researchers to formulate radically new research questions, increase visitor numbers, dramatically expand and diversify virtual access to our heritage, and bring clear economic, social and health benefits to communities across the UK. The innovation driven by the programme will maintain the UK's world leadership in digital humanities and set global standards in the field.

The Programme's main objectives are:

- to begin to dissolve barriers between different collections
- to open up collections to new cross-disciplinary and cross-collection lines of research
- to extend researcher and public access beyond the physical boundaries of their location
- to benefit a diverse range of audiences
- to be active and of benefit across the UK
- to provide clear evidence and exemplars that support enhanced funding going forward.

[35] https://tanc-ahrc.github.io/HeritagePIDs/

Aims of the Programme
The aim of the Programme is to begin to dissolve barriers between different collections, opening them up to new cross-disciplinary and cross-collection lines of research, and to extend researcher and public access beyond the physical boundaries of their location, thus directly addressing the issues related to accessibility beyond current metropolitan centres. The programme will extend across the UK including all the devolved nations, and will potentially have a global reach in terms of setting a standard for other countries building their own collections (with the long-term potential for inter-connection between the national collections).

This Programme will have a transformative impact on:

- Digital search and cataloguing tools, technologies and methodologies, and associated issues
- Research capability, by enabling search across collections to address cross-cutting research questions which will allow UK to maintain UK leadership in cross-disciplinary research
- The heritage sector as a whole, in terms of enhancing access for researchers, and for facilitating wider and better-informed public engagement.

There are two rounds of funding calls – the Foundation Projects and the Discovery Projects.[36]

[36] https://www.nationalcollection.org.uk/about