Project Name	**FREYA**
Project Title	**Connected Open Identifiers for Discovery, Access and Use of Research Resources**
EC Grant Agreement No	**777523**

D3.3 Prototypes of New PID Resources

Deliverable type	Demonstrator
Dissemination level	Public
Due date	29 February 2020
Authors	Robin Dasler (DataCite, https://orcid.org/0000-0002-4695-7874) Christine Ferguson (EMBL-EBI, https://orcid.org/0000-0002-9317-6819) Tina Dohna (PANGAEA, https://orcid.org/0000-0002-5948-0980) Uwe Schindler (PANGAEA, https://orcid.org/0000-0002-1900-4162) Frances Madden (British Library, https://orcid.org/0000-0002-5432-6116) Manuel Bernal Llinares (EMBL-EBI, https://orcid.org/0000-0002-7368-180X) Vasily Bunakov (STFC, https://orcid.org/0000-0003-3467-5690) Simon Lambert (STFC, https://orcid.org/0000-0001-9570-8121)
Abstract	This report describes the results of the prototyping implementations of new PID types and new PID services conducted by the FREYA partners. This work follows the previous two deliverables in this work package, which identified gaps in the PID landscape and determined feasibility of prototype implementation, respectively.
Status	Submitted to EC 6 March 2020 *Corrected version May 2020, fixing errors in internal cross-references to figures*

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

FREYA project summary

The FREYA project iteratively extends a robust environment for Persistent Identifiers (PIDs) into a core component of European and global research e-infrastructures. The resulting FREYA services will cover a wide range of resources in the research and innovation landscape and enhance the links between them so that they can be exploited in many disciplines and research processes. This will provide an essential building block of the European Open Science Cloud (EOSC). Moreover, the FREYA project will establish an open, sustainable, and trusted framework for collaborative self-governance of PIDs and services built on them.

The vision of FREYA is built on three key ideas: the **PID Graph**, **PID Forum** and **PID Commons**. The PID Graph connects and integrates PID systems to create an information map of relationships across PIDs that provides a basis for new services. The PID Forum is a stakeholder community, whose members collectively oversee the development and deployment of new PID types; it will be strongly linked to the Research Data Alliance (RDA). The sustainability of the PID infrastructure resulting from FREYA beyond the lifetime of the project itself is the concern of the PID Commons, defining the roles, responsibilities and structures for good self-governance based on consensual decision-making.

The FREYA project builds on the success of the preceding THOR project and involves twelve partner organisations from across the globe, representing PID infrastructure providers and developers, users of PIDs in a wide range of research fields, and publishers.

For more information, visit www.project-freya.eu or email info@project-freya.eu.

Executive summary

FREYA's work package on "New PID Types" (WP3) has been devoted to the exploration of new PID types and new services for existing PIDs. The first deliverable (D3.1, "Survey of Current PID Services Landscape") in the work package conducted an environmental scan of existing PID types, assessing their maturity and identifying gaps in the landscape. From these gaps, a subset of new PIDs and services was brought forward to the second deliverable (D3.2, "Requirements for Selected New PID Services"), which gathered user stories and determined the feasibility of developing prototypes to fulfil these user stories. The present deliverable, which is the third and final for the work package, describes the results of the FREYA partners' prototype implementations.

Ultimately, the FREYA partners prototyped four new PID types: PIDs for scientific instruments, PIDs for research facilities, PIDs for organisations, and PIDs for grants. They also prototyped additional services for PIDs, namely PID registration using metadata in landing pages and workflows for enhanced provenance metadata for digital collections.

Overall, no prototypes will be sunsetted after WP3 ends following the submission of this deliverable. Though some prototypes would require additional work or coordination to become fully production-ready services, they all made great strides over the course of the prototyping period. This work will be passed on to the work package responsible for integrating the PID Graph (WP4) as input for further enhancing the PID Graph and the vision of a world of interconnected PIDs.

Contents

1 Introduction

The work package on "New PID Types" (WP3) in the FREYA project has been largely concerned with the identification of needs for novel PID types and novel PID services, as well as exploration of the feasibility of implementing those PID types and services. In the wider context of FREYA, this work package serves as the experimental testbed, with the intention that it will pass on the results of its exploration to the work package responsible for integrating the PID Graph (WP4). To this end, WP3 began by identifying community gaps in regards to identifiers, compiling a wide list of needed PID types and services and assessing the current state of their readiness in the form of a maturity matrix (see D3.1, "Survey of Current PID Services Landscape"[1]). Based on this maturity evaluation, the FREYA partners identified a subset of PID types and services that would be taken forward as candidates for prototyping. In the next deliverable for this work package, the FREYA partners assembled a collection of user stories around these prototype candidates by seeking input from the wider PID community. The partners then determined the feasibility of developing these prototypical PID types and services within the constraints of the FREYA project.

In this final WP3 deliverable, the FREYA team is reporting on the prototypes that have been implemented and on what others can learn from our experiences. Because of the nature of prototyping, and the exploratory nature of this work package, not all prototypes will necessarily be taken forward as full-fledged production services, but the experimentation provides valuable insights into the benefits and pitfalls of developing such services and serves to inform the work of FREYA partners and others in expanding and improving the PID Graph. The lessons learned from this exploration, as well as the successful prototype services, can be taken up in WP4 as the FREYA project continues and carried forward as part of that work package's focus on integrating with the PID Graph.

As part of WP3's exploration, FREYA partners developed both new PID type prototypes and prototypes of enhanced services for existing PID types. The new PID types that were explored are:

- PIDs for scientific instruments (led by PANGAEA)
- PIDs for research facilities (led by STFC)
- PIDs for organisations (led by DataCite)
- PIDs for grants (led by EMBL-EBI, in partnership with Crossref)

The prototypes for enhanced services that were prototyped are:

- New workflows for PID registration (PANGAEA)
- JSON-LD metadata for identifiers.org (EMBL-EBI)
- Mechanisms of enhanced provenance information in digital collections (British Library)

In the following sections of this deliverable the new PID types are described one by one. For each new PID type the user story that was identified is presented, followed by the solution that was developed, as well as proposed next steps and lessons learned.

[1] https://doi.org/10.5281/zenodo.3554254

2 PIDs for scientific instruments

User story:
As a researcher, I would like to find data that was measured with the same instrument that I use in my research, to ensure that the measurements are comparable. This allows me to reuse data and increase the impact of my own data.

2.1 Solution

In the natural sciences, equipment such as seagoing vessels, platforms, buoys, sensors, sensor arrays or networks and other instrumentation are often central to data acquisition. While identifiers for vessels and platforms carrying instrumentation are relatively easily assigned, assigning identifiers for devices, instruments and sensors is more complex and these rarely bear any persistent identification other than inventory IDs in their owner's ledger. The Alfred-Wegener Institute (AWI), which coordinates German polar research, provides further solutions for equipment identification and accounting in research. It has recently (2015) initiated the '***Sensor Information System infrastructure***'[2] to support the flow of sensor observation to archives. They built a cost-effective and generic framework, the "Observations to Archive (O2A)", which complies with OGC standards, ensuring interoperability in an international context (e.g. SOS/SWE, WPS, WMS WFS,..). Each sensor is described following SensorML data model standards and data is fed to an SOS interface, so that the sensor can be monitored in real or near real-time. Scientists can register their instruments according to a set schema and receive a "handle" for the instrument description that they can then include in their data publication metadata. PANGAEA has started including these handles for instruments in dataset metadata. This is a first step and will make aggregating related data (based on instrument/sensor use) using machine-to-machine communication possible at PANGAEA.

[2] https://sensor.awi.de/

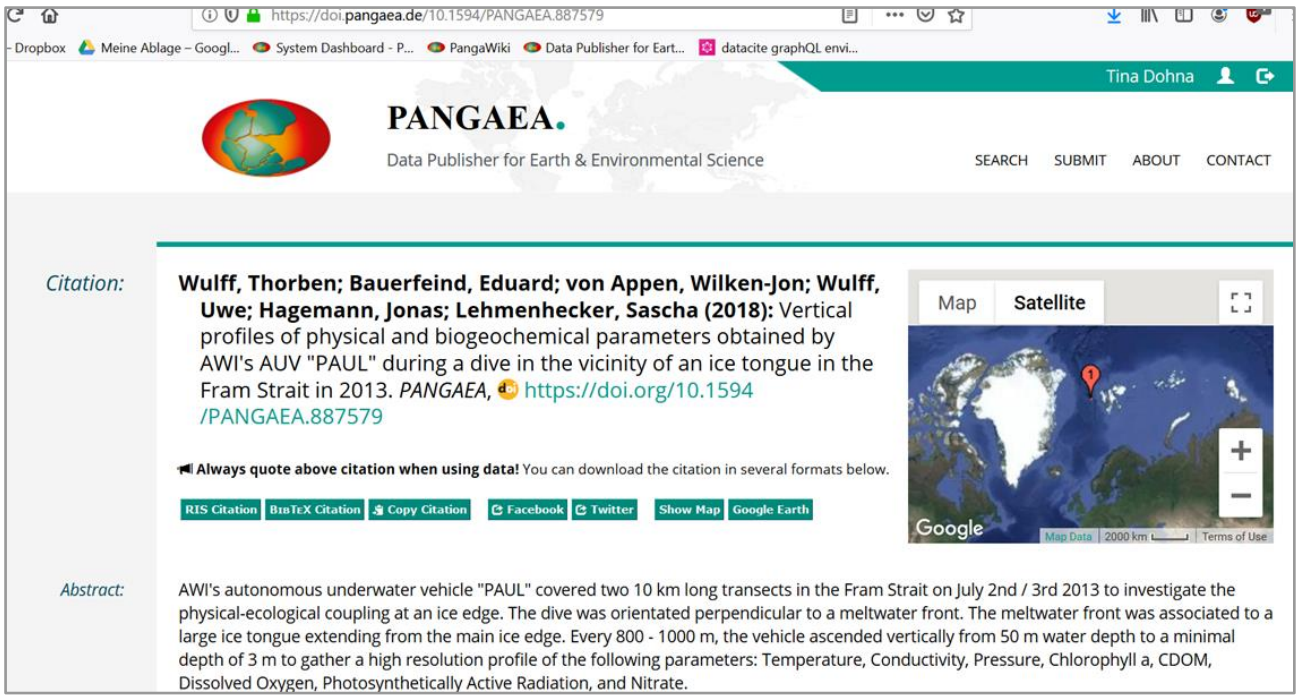


Figure 1 Example data set which has a PID for the instrumentation included in the data set metadata

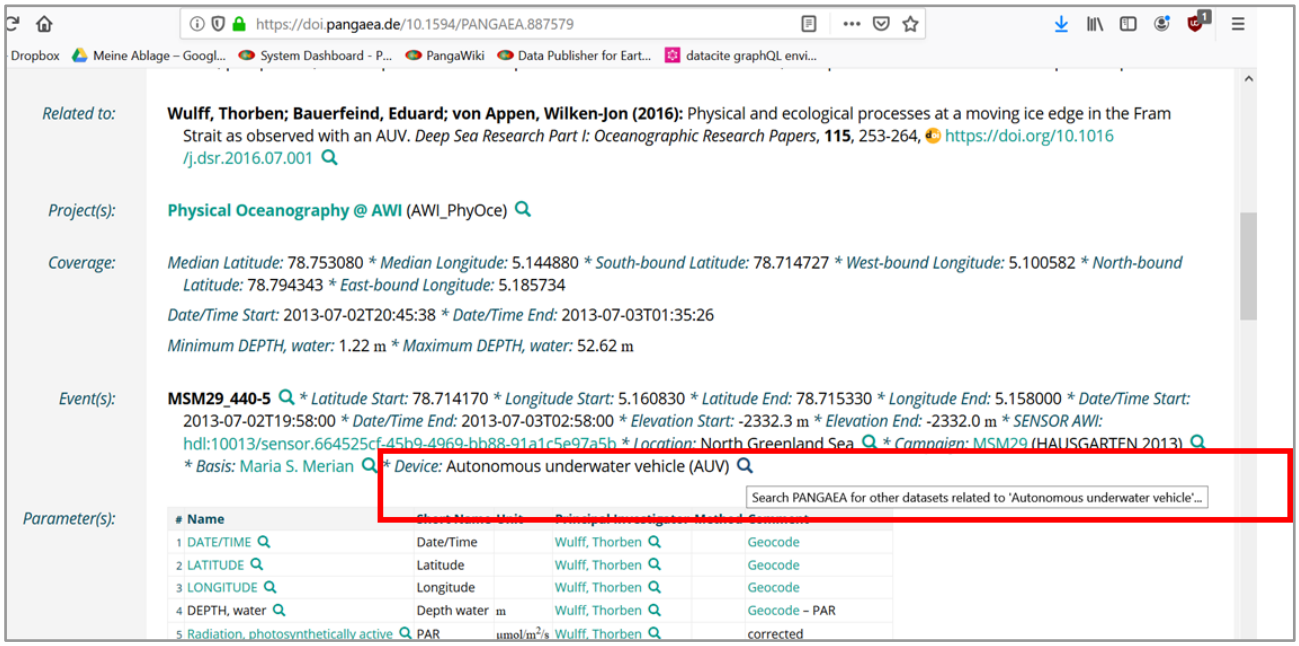


Figure 2 PANGAEA includes device types in the dataset metadata. Current work is mapping sensor.awi registry to PANGAEA device types to enable the functionality shown here ("search for similar datasets") to the more extensive instrument metadata recorded in the sensor registry.

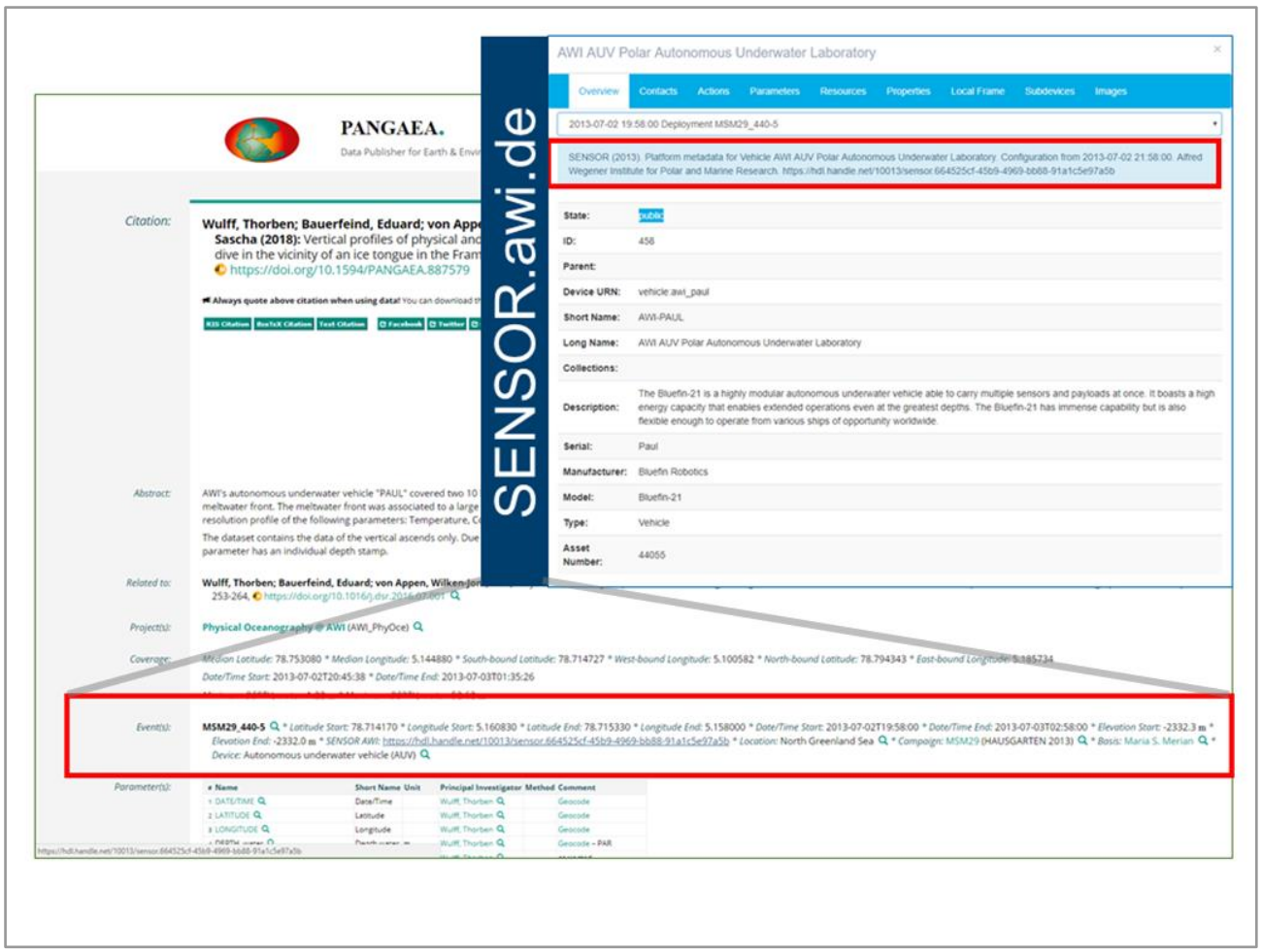


Figure 3

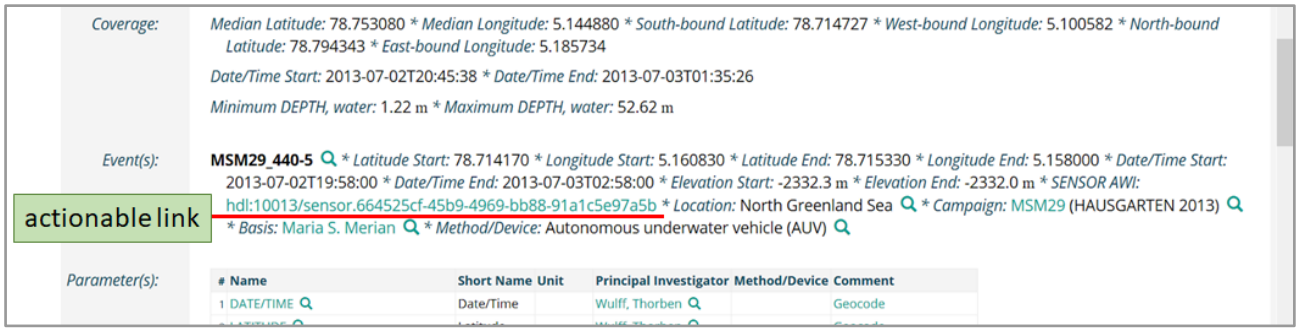


Figure 4 These two figures show the information provided to users when activating the handle for the instrumentation used in the research.

2.2 Next steps

Handles will be replaced by DOI registration once DataCite has accommodated the metadata schema from the RDA “Persistent Identification of Instruments” Working Group in their regular schema update by the end of 2020, thereby providing PIDs for instruments beyond an institutionally governed handle system. Standard vocabularies for the description of research-related entities need to be a strong focus of future work so that questions of interoperability between different research data platforms and communities, and also in the context of the EOSC, are addressed. The effort that goes into mapping disparately developed systems is immense and can be largely reduced if well-curated and complete vocabularies are available.

2.3 Lessons learned

Current bottlenecks for a measurable impact of the activity is the adoption of the new metadata schema by DataCite scheduled for the next schema update. This step will enable the registration of instruments with DOIs, adding an identifier that conforms to the stricter definition of a PID – compared to the handle used in this use case – and ensuring more widespread DOI registration. In addition, mapping the *sensor.awi* and PANGAEA vocabularies for devices/instruments/sensors has been extremely challenging, since legacy data is a very problematic aspect of this. The NERC vocabulary[3], used to this end, needs to be mapped onto the existing descriptions, and large gaps have been identified. PANGAEA is providing feedback to SeaDataNet to extend the vocabulary to include missing device types and models, so that complete mapping can be achieved and extended to other data archives.

Registration of devices includes filling in essential metadata for new devices, which has also presented a bottleneck in our experience. A change of culture is needed, so that scientists are more willing to invest time to provide this essential metadata component and thereby improve the discoverability and re-usability of their data.

[3] https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/

3 PIDs for research facilities

User story:
As a scientific facility's User Office, or as a funder supporting a facility, I would like to link facility awards (beamtime) with records in scientific databases, such as biomedical or crystallographic databases, so that I can determine the facility's impact.

3.1 Solution

PIDs for research facilities and those facilities' large-scale instruments can significantly contribute to research provenance in biomedicine and materials science, which are the two most prominent categories of facilities users (visitor scientists), as well as to the sensible integration of information sources curated with the consistent use of PIDs.

The FREYA deliverable D3.2 suggested that one candidate for prototyping at STFC could be the use of facility PIDs for linking the Diamond synchrotron[4] bibliography database with the Protein Data Bank (PDB) and EuropePMC. Additionally, an institutional repository with ISIS neutron and muon source[5] bibliographic records has been used, also the Inelastic Neutron Scattering database that contains spectral data from one of the ISIS beamlines. For the reference databases, the Cambridge Structural Database (CSD) has been examined, in addition to those initially planned.

Facilities present a special challenge for metadata modelling because of their three-fold nature as a funder, an instrument and an organisation, all of which are exploring their own possible PIDs. This makes the pursuit of a facilities PID an exercise in coordination and enhancement of existing initiatives and metadata schemas. Another implication of the user story that inspired this service prototype is that facility PIDs per se, or a selection of existing funder, instrument and organization PIDs that serve different aspects of a facility, are not going to be the core value of the service. The PIDs associated with facilities can rather be a tool for building a PIDs-rich knowledge graph that includes other types of persistent identifiers. In fact, other PID types (not those directly associated with facilities) can serve as tools for building the common graph, too; one example of this is bibliographic records (bearing DOIs) that are often collected by facilities as proof of their research impact. Building a PID-rich knowledge graph around facilities research is a multilateral and multidirectional exercise, and this graph rather than a specific PID type has a real potential to underpin a new service that can be of value to various stakeholders within facilities and beyond their organizational walls.

Bearing these wider considerations in mind, the structure of the Diamond database was assessed and the PDB API was explored. Making connections between the Diamond database and the PDB proved possible yet requires more development effort to integrate the PDB identifiers in a common graph; implementation of this integration will be aligned with the remaining WP4 effort. For EuropePMC, over 2000 potential matches with the Diamond database have been discovered using PubMed IDs; the implementation of the actual connections will be aligned with the WP4 effort. Additionally, another resource for integration—

[4] http://www.diamond.ac.uk/
[5] https://www.isis.stfc.ac.uk/

Cambridge Structural Database (CSD)—was explored, and two mappings were produced: between Diamond and CSD records (about 700 publications DOIs matched to 3500 structures in the CSD) and between ISIS and CSD records (about 300 publications DOIs matched to 1300 structures in the CSD). The implementation of the actual connections across the three sources will be aligned with the WP4 effort.

The technology foundation for the service is the graph database that integrates metadata from a few publication repositories and data repositories within STFC, enriched with references to the external well-curated databases beyond the organization walls. The prototype currently uses STFC firewalled virtual machines as an infrastructure, and requires more development effort to mature it to a publicly available beta-version. The remaining effort in WP4 focussed on integration can be used to make more progress within the FREYA lifespan, yet STFC will have to extend this effort with additional resources to make the sound technological foundation of the service. The service will require a clear definition of a governance model, too; the remaining effort in WP6 can contribute to this, yet again the FREYA effort in this respect should be matched with the STFC own resource. The discussions about the technological and governance aspects of the service are ongoing and will be intensified towards the end of 2020, to fit them into the actual planning lifecycle within the organization.

The service has potential value for external stakeholders beyond the walls of an organization that operates research facilities. One example of this is a contribution to provenance records in reference databases, which WP3 has used to explore the opportunity for improved research provenance. Identification of facilities that contributed to the data records in Protein Data Bank, EuropePMC and Cambridge Structural Database could improve data provenance in them, in addition to naturally captured provenance of publications (attributed to facilities) in STFC publication repositories. The aforementioned numbers of the record matches found imply that we can aim for a contribution of around 7000 provenance records across the EuropePMC and Cambridge Structural Database, with a potential for the Protein Data Bank to be further explored. Another contribution to provenance has been the improved attribution of doctoral theses in the British Library EThOS repository in respect to STFC that either sponsored the PhD research directly (in a monetary form) or supported the PhDs with grants-in-kind (facility time). The provenance of about 600 EThOS records can be improved this way; the actual implementation of the metadata with improved provenance will be completed once EThOS is migrated to a new platform in 2020/21.

The question of granularity of the involved agents and processes is important for modelling provenance, so a certain effort has been devoted to modelling not only the facility as a whole but also its large-scale instruments (beamlines). The DataCite Metadata Schema was assessed to see if it suits the requirements of metadata for facilities and facility instrument PIDs. The use of the DataCite Metadata Schema seems reasonable and can be pursued with the proper “buy-in” from facilities who should take ownership of their PIDs in order to sustain them. Two templates for facilities large-scale instruments (beamlines) were produced using the DataCite schema elements: one for a Diamond synchrotron beamline and another for an ISIS neutron source beamline. The templates are used in an ongoing discussion with the respective facilities stakeholders, to ensure the metadata in them suit the stakeholders’ needs and can be continuously supported in the actual state by the facilities themselves.

Another ongoing development using a similar approach has been using PIDs for augmenting the Inelastic Neutron Scattering (INS) database, which is a database resulting from research on the ISIS neutron and muon source in STFC, primarily focussed on the investigation of materials and novel chemical substances. This should rely again on the long-term support of facility beamline PIDs by the facility who should “own” and sustain their PIDs, which is mainly a question of new best practices for the database records curation. The long-term sustainability can be achieved by the incorporation of the resulting INS PID graph into the STFC Open Science Portal, in addition to the aforementioned integrations of the STFC publication repositories with Protein Data Bank, EuropePMC, Cambridge Structural Database and the British Library EThOS service.

3.2 Next steps

The prototyping work is currently at a moderate level of maturity; the plan is to carry this work forward into WP4 as indicated above in Section 3.1 and mature the technology aspect of the prototype as much as possible during the remainder of FREYA. The aim is that this work can be incorporated into an Open Science Portal for STFC, with the goal to further develop and support it long-term using STFC own resources. The Portal will require a reasonable governance model, too and the remaining STFC effort in WP6 will contribute to the model definition. The work on provenance mentioned in Section 3.1 leads to an open question of whether any existing provenance model can be consistently applied to the facilities research case. This research opportunity for the proper modelling of provenance will be pursued opportunistically and with less priority than the service development.

3.3 Lessons learned

As with other implementations in this report, the coordination between other entities and projects that have their own timeline external to the implementation at hand can cause a natural delay to implementation. As an example, the organizational aspect of facilities can be addressed by the emerging PID services like ROR; for the funding aspect, the Crossref directory may suit and for the instrumental aspect, the DataCite metadata schema already presents some modelling opportunities which will be further improved with the planned schema updates. Yet these pieces of a comprehensive facility metadata associated with the respective PIDs are currently at different levels of maturity, which makes the facility metadata design challenging. Another lesson learned is that a service prototype that emerged from a particular use case can naturally lead to the better vision of a service that is really worth sustaining and where PIDs, their metadata and their interconnections are considered a new information infrastructure that can serve various use cases rather than the one initially considered. So the service development is not entirely one direction from an initial concept to implementation, but can be “back-propagated” from the considerations of what is really worth sustaining, what can deliver real value for the service immediate and prospective users.

4 PIDs for organisations

User story:
As a university administrator, I want to get a list of all datasets and software published by our researchers, so that I can get a comprehensive view of our research outputs.

4.1 Solution

FREYA partner DataCite is collaborating on building ROR[6], the Research Organisation Registry, along with several partners external to FREYA, namely Crossref, California Digital Library, and Digital Science. This group released a "minimum viable registry" for ROR in January 2019. This registry was initially based on data ingested from GRID, with the intention that this data will later be expanded by contributions from organisations themselves, mediated by human curation.

While the general collaboration on ROR exists outside the time constraints of the FREYA project, the work to incorporate ROR into DataCite services was carried out as part of the prototyping efforts of FREYA. These efforts have resulted in an initial offering of production-level services that make ROR IDs available to every DataCite member.

This work began with incorporating ROR IDs into the DataCite DOI Fabrica platform, which is the primary web platform DataCite members use to create and manage DOIs and DOI metadata for use in their own repositories and journals. The ROR ID was added as an available name identifier for an organisational name (as in the case of an item authored by an organisation), and the Fabrica platform made use of the ROR registry to look up valid name strings for inputting affiliation information. Further, DataCite staff took on the task of adding ROR IDs to the member profile information of all DataCite members. This information is currently only visible to members, but it can be used behind the scenes to more accurately link members to other services.

[6] https://ror.org

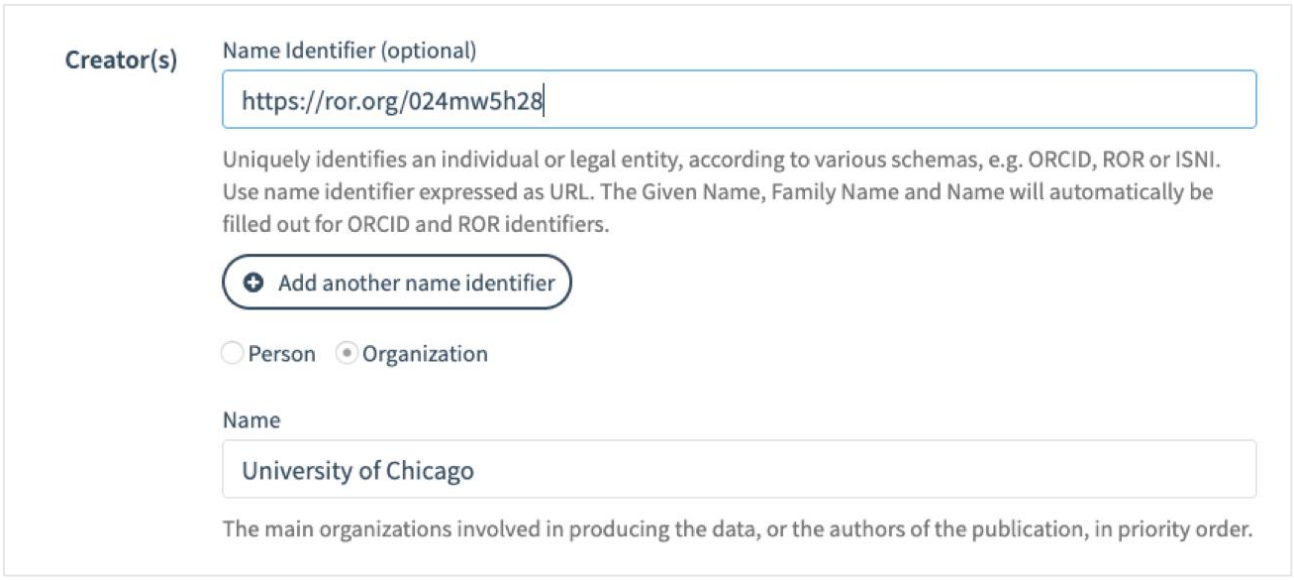


Figure 5 In the DataCite DOI Fabrica platform, pasting a ROR ID into the Name Identifier field will look up the relevant organisation in the ROR registry and automatically populate the appropriate name information

This initial work described above, completed prior to the FREYA midterm review, was necessarily limited by the fact that the DataCite Metadata Schema did not yet accommodate ROR. The work to update the Schema and to update the corresponding functionality in both the DataCite DOI Fabrica platform and in DataCite APIs has been undertaken since the midterm review, resulting in an updated DataCite Metadata Schema 4.3 released in August 2019 and resulting in additional options for including ROR in a DOI record created via DataCite services.

The most important addition to DataCite Metadata Schema 4.3 was the creation of a field for an affiliation identifier, which was not previously part of the schema. This allows DataCite members to include a ROR or other organisational identifiers associated with an author's affiliation. This field is repeatable and an affiliation can be expressed for each author of the item receiving the DOI. In a similar vein, DataCite Metadata Schema 4.3 also saw the addition of ROR as a possible option for identifying a funder.

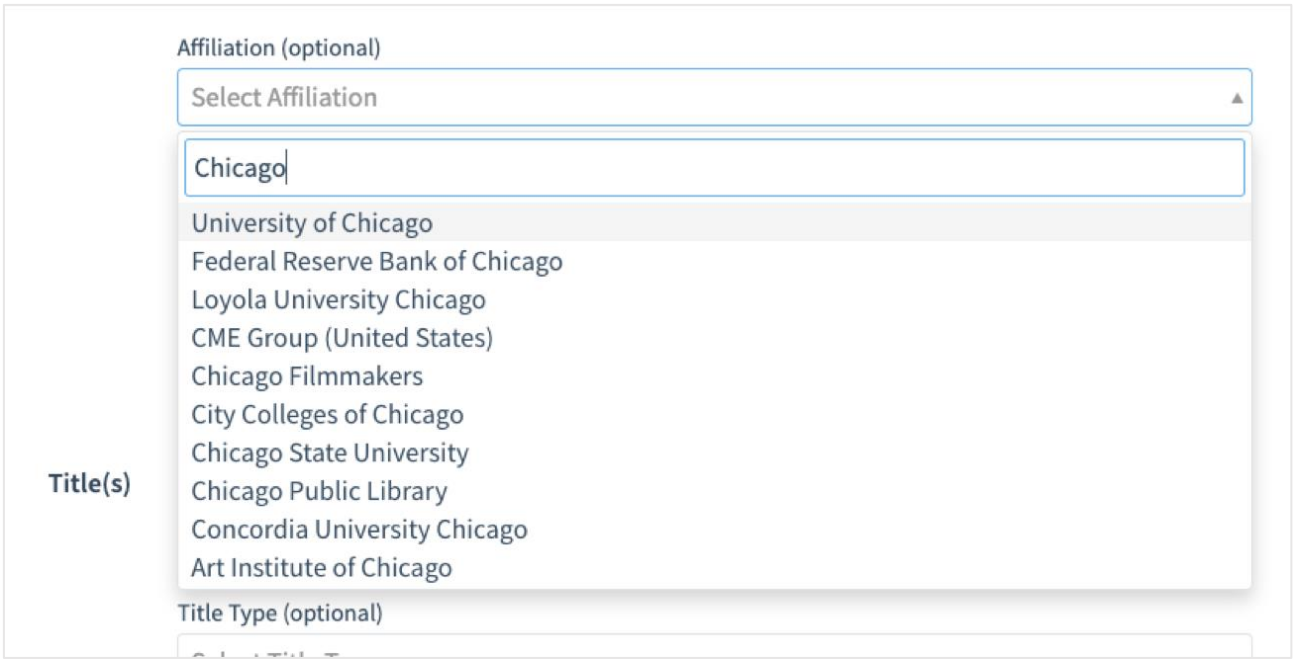


Figure 6 In the DataCite DOI Fabrica platform, searching for an organisation in the Affiliation field looks up the relevant entry in the ROR registry

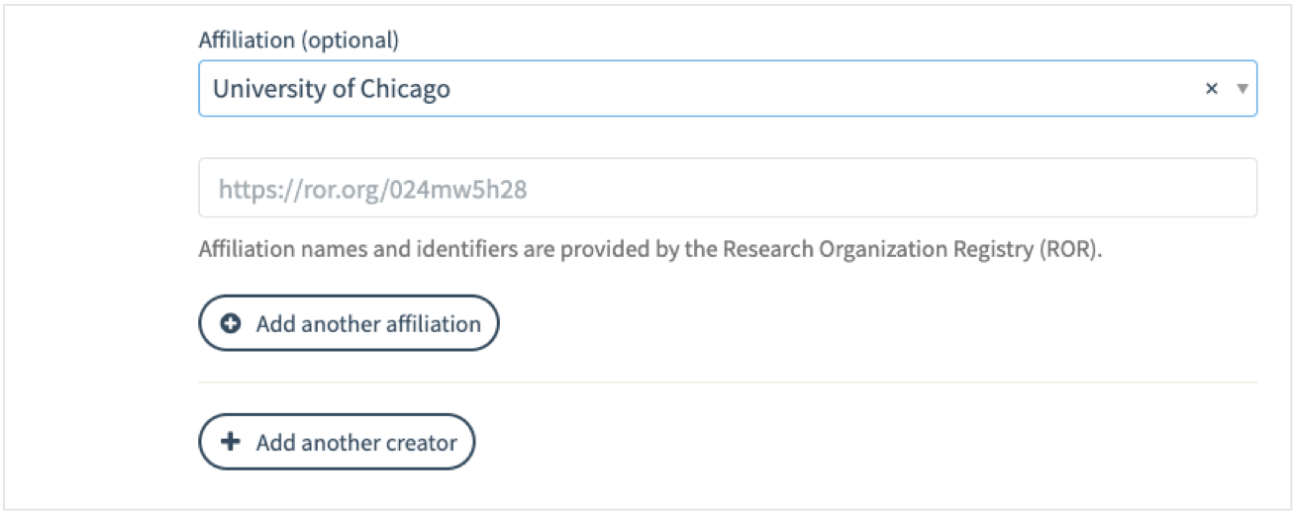


Figure 7 An example of an automatically populated ROR ID in the Affiliation field in the DataCite DOI Fabrica platform

With these metadata changes, all DataCite members are now free to include ROR IDs in the metadata that is submitted to DataCite. These members include several FREYA partners, as well as many universities and national library repositories, so the reach is potentially quite broad. One early adopter of ROR is the Dryad repository, who made a major effort to add ROR IDs to their back catalogue of datasets. As of the time of this writing, there were 44,628 DOI records across all of DataCite that contained an affiliation identifier.

It should be noted that much of the work to incorporate support for ROR IDs into the DataCite Metadata Schema and into the DOI Fabrica platform was completed prior to the release of FREYA deliverable D4.4, "Organizational IDs in Practice"[7], as the completion of this work was necessary to enable the other FREYA partners to implement their own organisational ID solutions. As such, a brief description of the state of ROR was included in D4.4, but the actual work to implement ROR IDs across DataCite services was undertaken as part of WP3, and is therefore presented here.

4.2 Next steps

The functionality to include ROR IDs as affiliation identifiers, funder identifiers, or name identifiers for an organisational creator is now an established part of DataCite's normal operational services. DataCite members have begun to use this functionality, but thus far the number of DOI records containing ROR IDs is a small percentage of the nearly 20 million DOIs registered with DataCite. Increased outreach efforts are necessary to encourage DataCite members to update their DOI metadata to include ROR IDs.

ROR will continue to develop beyond the scope and lifetime of FREYA. The most pressing goals for ROR are:

- to develop curation procedures and policies so that organisations can participate in the curation of their own data;
- to ensure the financial sustainability of the service while keeping the data in the registry free for public use; and
- to promote adoption and integration with further services.

Participation in FREYA has helped with the third goal by providing a ready cohort of early adopters, in the form of the FREYA partners, and by promoting ROR through the FREYA ambassador network, thus spreading the potential for future adoption. In addition to FREYA partners, implementations involving ROR

[7] https://doi.org/10.5281/zenodo.3606059

are under development at Dryad[8], Altum[9], Cobaltmetrics[10], Rescognito[11], Data Salon[12], Imperial College London[13], and Open Access Button[14].

4.3 Lessons learned

Implementing PIDs for organisations in DataCite services, as with the other implementations described in this report, is a reminder of the number of "moving parts" involved in this type of enterprise. Even after the PID itself is made ready for use by the authority that has designed it, there is still significant effort required to account for its use in standardised metadata schemas, to plan for its inclusion in united services, and to develop the user interfaces to allow researchers and data managers to use it. In the case of organisation PIDs at DataCite, we are already involved in every stage of the pipeline, from PID design to UI implementation, so the barriers are manageable on our own timetable. For others, successful implementation of new PIDs may require coordination with multiple external entities and reliance on timetables outside of their control.

[8] https://datadryad.org/
[9] https://www.altum.com/
[10] https://cobaltmetrics.com/
[11] https://rescognito.com/
[12] https://www.datasalon.com/
[13] https://www.imperial.ac.uk/
[14] https://openaccessbutton.org/

5 PIDs for research grants

User stories:

As a funder, I would like to identify published outputs of a grant we've awarded so that we can assess the impact of the grant. I would also like to know which researchers are related to a grant, for grant management purposes.

As a researcher, I wish to link published output to a grant, to acknowledge a funder and thereby satisfy a funder's mandate. In literature searches where a grant is linked to a study, I would like to discover the details of research to be funded by that grant, as well as any other publications stemming from the grant. This will give me insights into the status of the specific research being funded.

5.1 Solution

To begin to satisfy the above user stories, a global grant identifier system[15] is needed, such as that being developed by the Wellcome Trust and FREYA partner Crossref. The benefits of such a system would be that the "identification of grant-specific research outputs [is made] more accurate, whilst simultaneously reducing the burden on the researcher" by automating the process.

In order to implement a global grant identifier, two things are needed. First, all new grants must be assigned a unique ID. For grants, it was agreed that the unique IDs will be Digital Object Identifiers (DOIs). Second, every DOI must resolve to a publicly accessible web site (e.g. http://europepmc.org/grantfinder), where information about that grant is disclosed.

Europe PMC at FREYA partner EMBL-EBI partnered with Crossref, Wellcome, and PLOS on this initiative. This report focuses on implementations by Europe PMC.

[15] https://www.crossref.org/blog/wellcome-explains-the-benefits-of-developing-an-open-and-global-grant-identifier/

Figure 8 Metadata fields that Europe PMC provides to Crossref when registering a global grant ID

Country	RORID	Role	Funding Type	Contributed Amount	Contributed Currency	Funding Percentage
GB		lead_investigator	award	200000	GBP	20
GB	ror.org/0220mzb33	investigator	fellowship			
AT		co-lead_investigator	grant			

Figure 9 Additional information is now required by Crossref in order to mint DOIs for grants. What's new is the 'role' and 'funding type'

Europe PMC obtains grant data from its funders, which is then stored in Europe PMC's GRIST (GRant Information SysTem) database. For the grants DOI project, the GRIST database needs to be expanded to allow for collection of additional grant information. Figure 9 shows the newly required information (fields) that funders are requested to provide – these are listed in columns in this spreadsheet. The items in blue are required by Crossref when creating a DOI. The items in orange are optional.

FundingTypeID	Name	Description	DateCreated
1	award	Award	20/11/2019 13:38:08
2	contract	Contract	20/11/2019 13:38:08
3	crowdfunding	Crowdfunding	20/11/2019 13:38:08
4	endowment	Endowment	20/11/2019 13:38:08
5	equipment	Equipment	20/11/2019 13:38:08
6	facilities	Facilities	20/11/2019 13:38:08
7	fellowship	Fellowship	20/11/2019 13:38:08
8	grant	Grant	20/11/2019 13:38:08
9	loan	Loan	20/11/2019 13:38:08
10	prize	Prize	20/11/2019 13:38:08
11	salary-award	Salary award	20/11/2019 13:38:08
12	secondment	Secondment	20/11/2019 13:38:08
13	seed-funding	Seed funding	20/11/2019 13:38:08
14	training-grant	Training grant	20/11/2019 13:38:08
15	other	Other	20/11/2019 13:38:08

Figure 10 Screenshot of one of the new "tables" added to the GRIST database. This particular table stores the grant funding types provided by Crossref.

For this project EuropePMC has collaborated with Wellcome to gather the data for their 2019 grants. Europe PMC will submit this grant data to Crossref, which will allow Wellcome, now a Crossref member, to have DOIs minted for their grants. Figure 11 gives an example of a DOI for a Wellcome-funded grant – showing the prefix that will be assigned to all Wellcome-funded grants in red.

Wellcome funder prefix

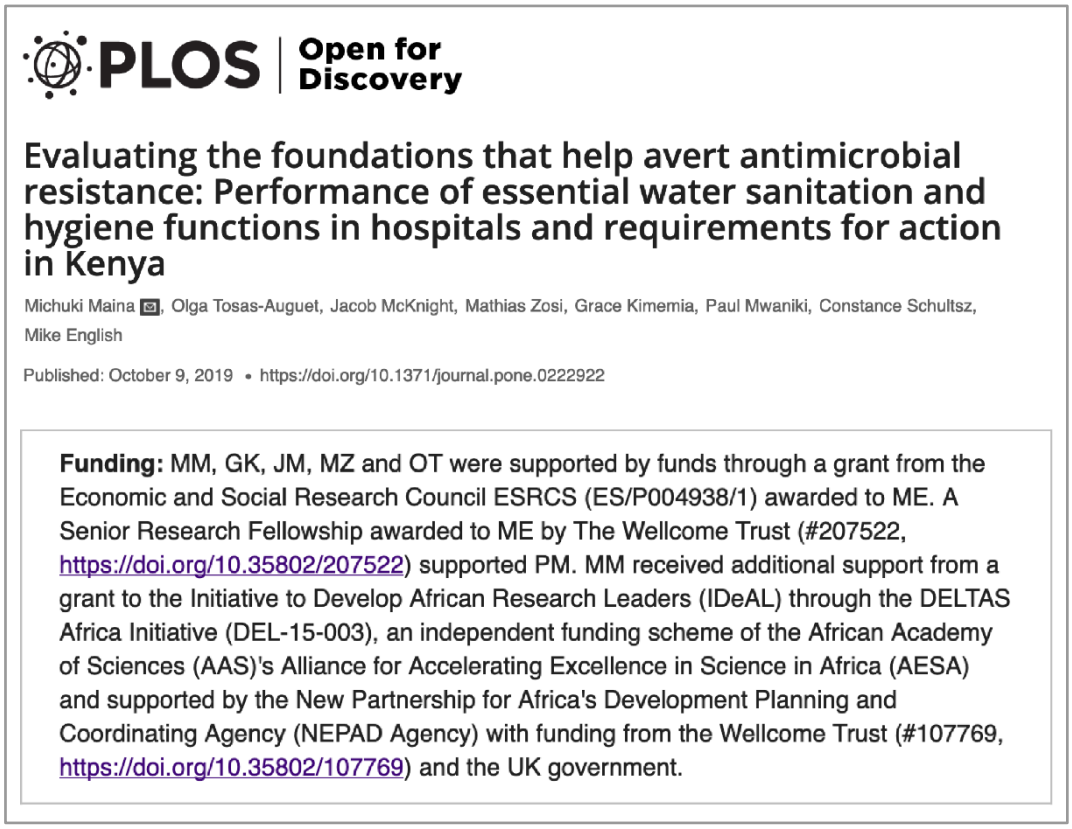
https://doi.org/10.35802/107769

Figure 11 Example of a DOI for a Wellcome-funded grant

What we hope to see is that authors disclose funding information to publishers on submission, which is published and can also be passed on programmatically.

In this proof of principle pilot, the journal PLOS ONE[16] has coordinated with Wellcome-funded authors to include newly registered global grant IDs in the metadata of the publication. This means that the readers can now seamlessly navigate from the article to the grant record and examine the support provided by Wellcome for this particular study.

[16] https://journals.plos.org/plosone/



PLOS Open for Discovery

Evaluating the foundations that help avert antimicrobial resistance: Performance of essential water sanitation and hygiene functions in hospitals and requirements for action in Kenya

Michuki Maina, Olga Tosas-Auguet, Jacob McKnight, Mathias Zosi, Grace Kimemia, Paul Mwaniki, Constance Schultsz, Mike English

Published: October 9, 2019 • https://doi.org/10.1371/journal.pone.0222922

Funding: MM, GK, JM, MZ and OT were supported by funds through a grant from the Economic and Social Research Council ESRCS (ES/P004938/1) awarded to ME. A Senior Research Fellowship awarded to ME by The Wellcome Trust (#207522, https://doi.org/10.35802/207522) supported PM. MM received additional support from a grant to the Initiative to Develop African Research Leaders (IDeAL) through the DELTAS Africa Initiative (DEL-15-003), an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (#107769, https://doi.org/10.35802/107769) and the UK government.

Figure 12 Global grant IDs for Wellcome grants featured in the Funding section of a PLOS ONE publication

Clicking on the first grant DOI mentioned in the published article takes the reader to a landing page for that Wellcome-funded grant in the Grist database (Figure 13).

Figure 13 : A landing page for a Wellcome-funded grant in the GRIST database

Notably, all of this grant-associated metadata is freely available not only on Europe PMC's website but also programmatically, through the public GRIST API[17]. The newly created global grant ID along with the local grant number have been incorporated into the API response. The grant IDs and associated metadata will be available via Crossref's APIs[18] later in 2020. DOIs for grants have to date (Jan 2020) been registered on behalf of Wellcome for 237 grants awarded in 2019.

[17] https://europepmc.org/GristAPI
[18] https://www.crossref.org/services/metadata-delivery/

5.2 Next steps

Grant IDs will be assigned retrospectively to Wellcome grants awarded and registered in Europe PMC's GRIST database from years prior to 2019. This will encompass approximately 13,500 Wellcome grants currently available in the GRIST database.

As a long-term aim, the adoption of global grant IDs will allow us to create a more interlinked PID Graph. As Europe PMC hosts data for both publications and grant awards, we are well positioned to link publication DOIs with DOIs for grants, supporting better tracking of the research funding impact. We hope that by implementing global grant IDs, grant data can be easily collected on submission by publishers and repositories and automatically fed into researcher assessment platforms, thereby simplifying researchers' workflows.

5.3 Lessons learned

This pilot implementation by Europe PMC of DOIs for Wellcome Trust grants that are indexed in Europe PMC's GRIST database serves as a demonstrator for further implementations. The requirements established are as follows:

Funders:

- will require membership of Crossref in order to register DOIs for grants. Crossref provides a metadata schema for grant information. What is needed is a means to provide the metadata to Crossref. This is a service currently provided by Europe PMC for its 29 funders.
- will require landing pages to feature the metadata associated with each grant DOI. Currently EuropePMC provides landing pages within the GRIST database for its 29 funders, and makes this information available to users via its grantfinder search interface[19] and programmatically via a public GRIST API[20].

Researchers:

- will need to obtain grant DOIs for any grants they have been awarded and include this in any funding statement associated with a research output such as a publication or dataset.

Publishers:

- will need to build into their publication workflows a request to authors to include grant DOIs. The presence of the grant DOI in the metadata will be sufficient to identify the funder and the specific grant as this is required in the metadata connected to the grant ID.
- will submit grant DOIs to Crossref as a piece of their existing article metadata when they register content with Crossref. (Metadata guidelines to be released mid-2020.) It should be noted that many publishers already collect funder-internal award numbers, albeit just as an open text field.
- should also publish the grant ID so that it is easy for anyone reading the paper to find information on how the research was funded.

Research Managers and repositories and other infrastructures such as EOSC:

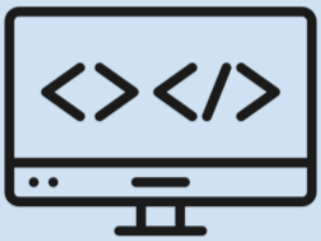
- will need to build into their workflows and systems/platforms a request to researchers for grant DOIs, or programmatically ingest this information with the other metadata on the research outputs they host. It should be noted that it is possible that grant DOIs can be linked to funder publication and data sharing policies to help ensure requirements are met in a more automated way.

[19] http://europepmc.org/grantfinder
[20] https://europepmc.org/GristAPI

These steps in turn will lead to sufficiently widespread grant PID information that could be incorporated in a research graph.

The Europe PMC pilot implementation has provided a template for other funders to use when implementing Grant DOIs. This includes the 29 Europe PMC funders, but Crossref is working with a large range of European and international funders (such as the Japan Science and Technology Agency and OSTI/DOE in the US) on their own implementation of these identifiers and workflows.

6 New workflows for PID registration



User story:
As a new data center I would like to avoid having to produce my own metadata in XML format for DOI registrations of dataset publications. I would like to achieve the registration with the metadata already embedded in schema.org/JSON-LD format in my dataset landing pages.

6.1 Solution

The current workflow for registrations of new DOI names is a two step process. As a first step, the data center calls a webservice to mint a DOI name. This makes the DOI resolvable and usable on the web. The API call takes the DOI name and the URL of the landing page. In a second step, data centers like PANGAEA currently have to create an XML metadata document in the proprietary DataCite metadata format. Those additional steps require mapping of internal metadata schemes to the DataCite Metadata Schema.

As part of a FREYA WP2 effort, DataCite implemented a way to process a DOI registration in a single step, opening the flexibility to use alternative metadata formats like Schema.org/JSON-LD. Many data centers already include metadata using the Schema.org standard into their dataset landing pages (hidden to end users, but readable for machines), because internet search engines like Google and its Google Dataset Search use this format to populate their search indexes. The new workflow combines the minting of DOIs with a landing page URL (to allow resolving the DOI) with extraction of the metadata embedded into the landing page in Schema.org format. The second step to separately upload DataCite Metadata in XML format can be omitted.

Figure 14 shows an example dataset's landing page. Invisible to the user, the source code of the HTML landing page also contains the whole dataset metadata in JSON-LD format using the Schema.org standard. The source code was made visible in the red box. Figure 15 shows the DOI minting and metadata submission process (using DataCite's DOI Fabrica web interface). Instead of uploading the metadata in the proprietary DataCite format, the URL to the landing page is given two times: (1) As target URL for the redirect installed on doi.org; and (2) instead of the metadata. When doing this, the webservice behind DOI Fabrica automatically loads the landing page and extracts the previously shown JSON-LD metadata. Figure 16 finally shows the imported metadata in DOI Fabrica.

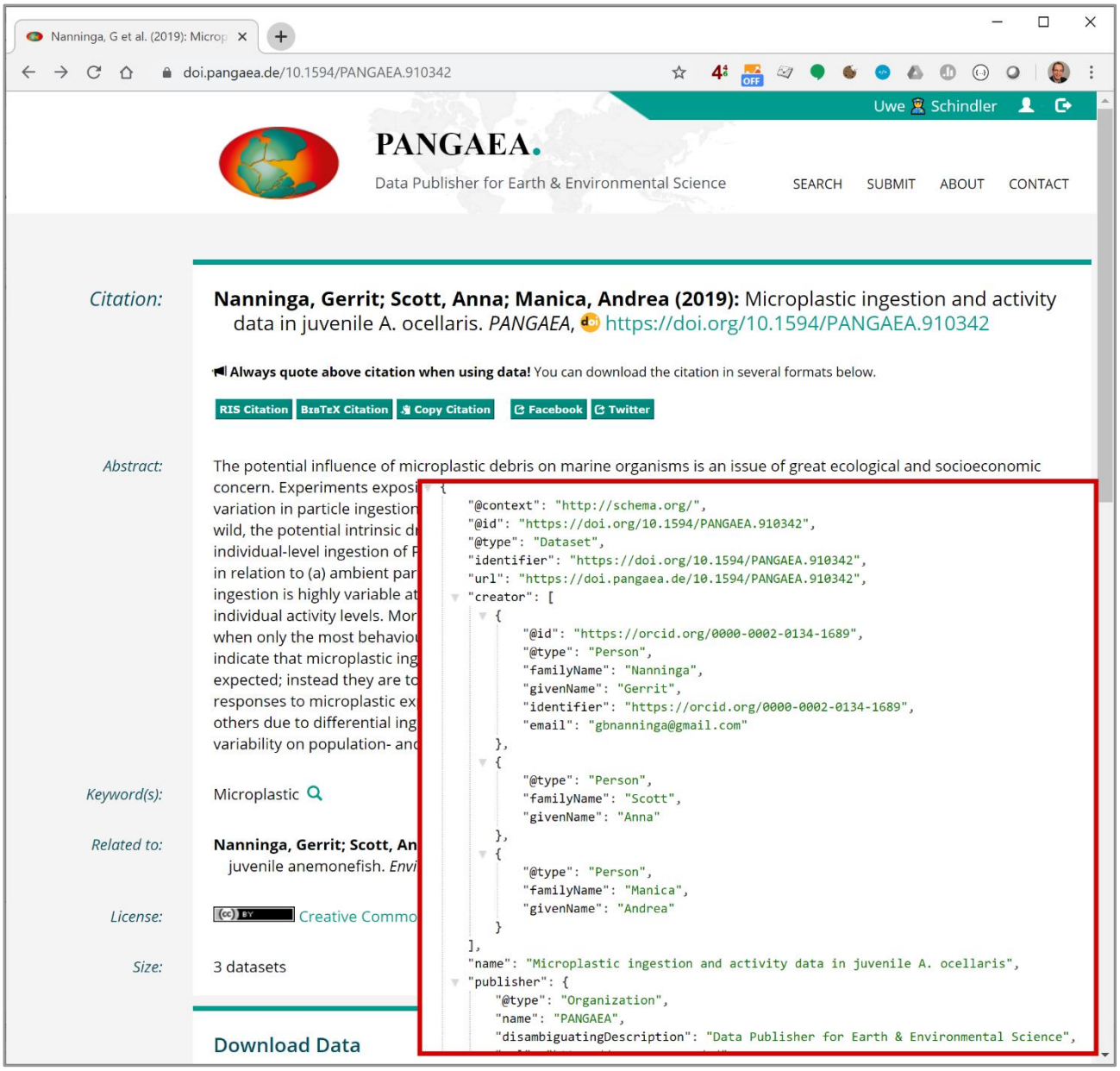


Figure 14 Example of a PANGAEA dataset with JSON-LD metadata in Schema.org format (red rectangle) embedded in HTML source code

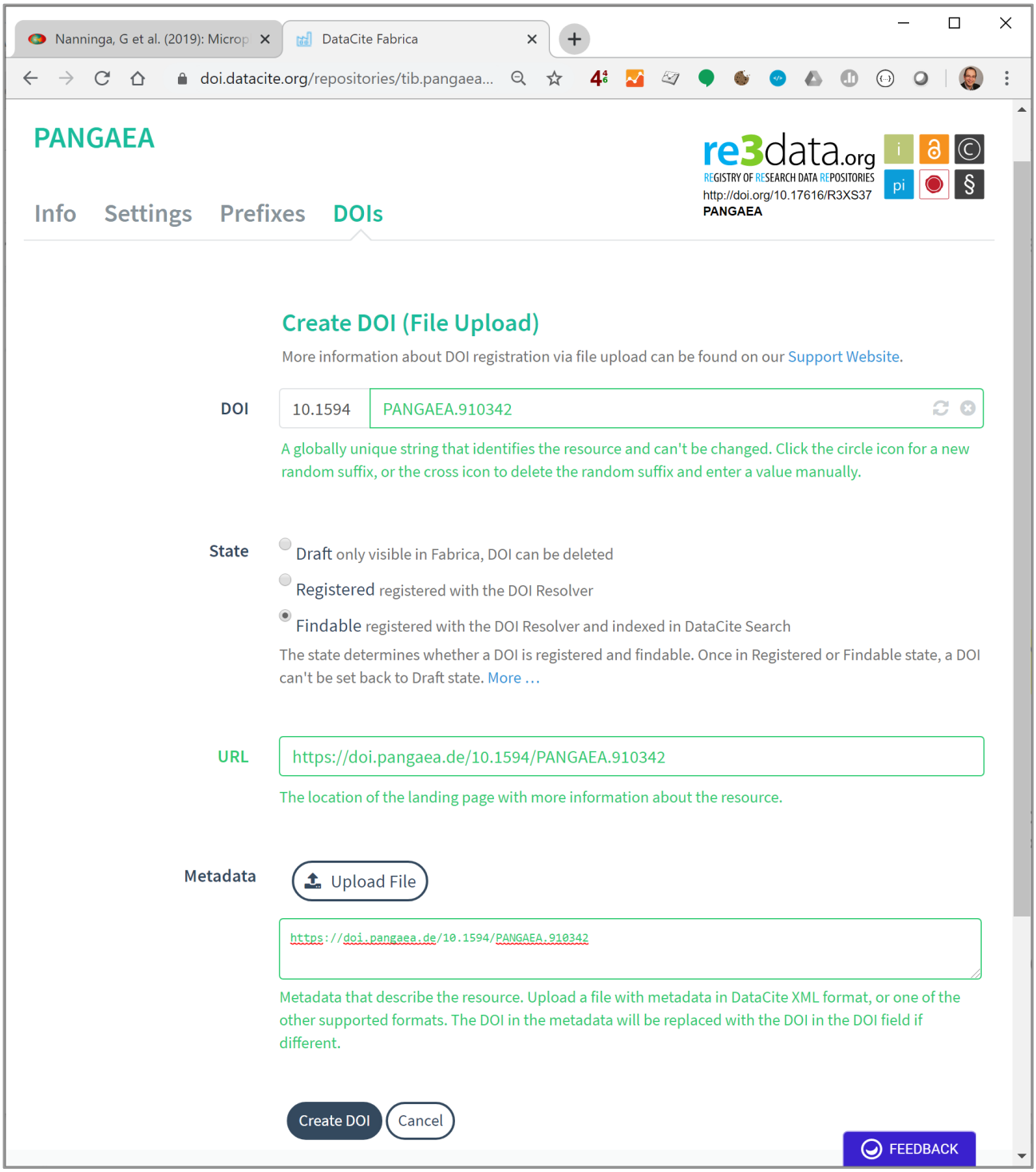


Figure 15 Specifying the landing page in DataCite DOI Fabrica

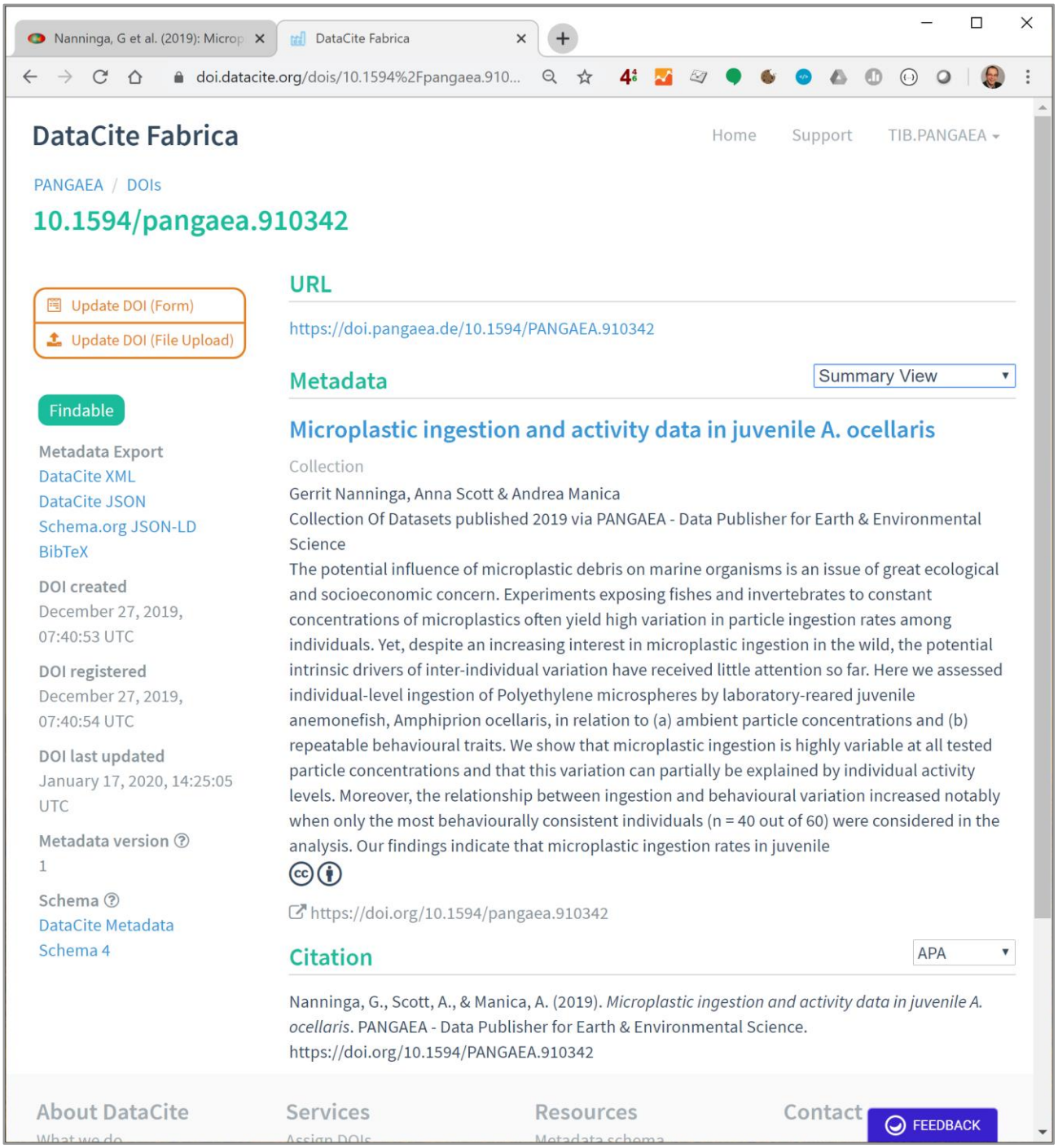


Figure 16 Example dataset with minted DOI and metadata extracted from the JSON-LD Schema.ORG metadata in DOI Fabrica

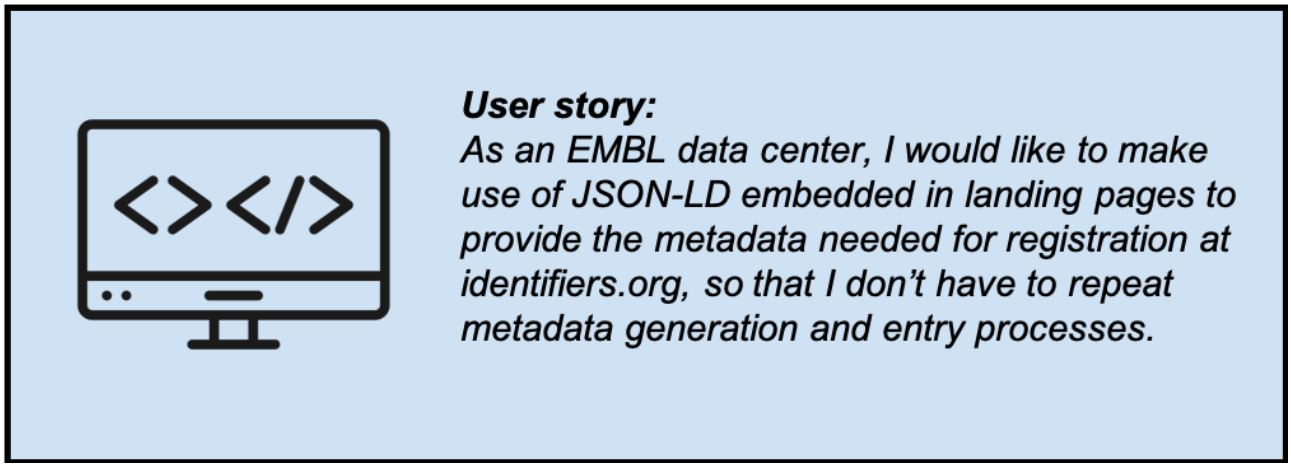
6.2 Next steps

PANGAEA found out that some of the more specific metadata elements do not map perfectly, so additional work is needed to extend the Schema.org metadata on their landing pages. At the time of the tests, for example, funding information was not yet included.

6.3 Lessons learned

PANGAEA tested the integration in their productive infrastructure by registering some DOIs using the new API calls / DataCite DOI Fabrica user interface and compared the results. Although Schema.org metadata and DataCite metadata differ in how they describe the data, it is still possible to extract most of the relevant information to fully describe a registered DOI from the landing pages. This makes adoption for new data centers much easier; especially research infrastructures such as EOSC will only need to build Schema.org metadata into their research output web pages allowing metadata extraction by PID providers like DataCite out of the box. In terms of making metadata more easily accessible this way, web search engines like Google Dataset Search can also pick up this metadata and disseminate it.

7 Identifiers.org and JSON-LD metadata



7.1 Solution

The Identifiers.org system is a central infrastructure for findable, accessible, interoperable and re-usable (FAIR) data. It provides a range of services to generate, resolve and validate persistent Compact Identifiers to promote the citability (see Figure 17) of individual data providers and integration with e-infrastructures[21,22].

The Identifiers.org registry contains hundreds of manually curated, high quality data collections, with each assigned a unique prefix. A combination of the prefix and a locally assigned database identifier (accession) forms a Compact Identifier, [prefix]:[accession]. For example, pdb:2gc4, GO:0006915, etc.

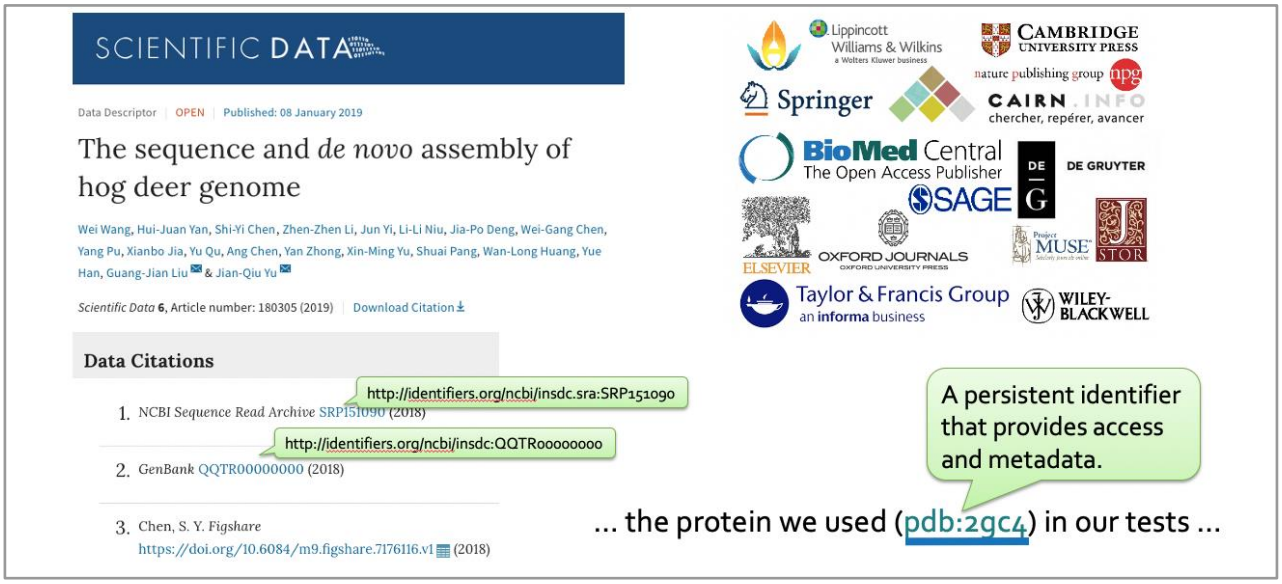


Figure 17 Example of Compact Identifiers for in-line data citation

[21] Sarala M. Wimalaratne et al. Uniform resolution of compact identifiers for biomedical data. Sci. Data 5:180029 doi: 10.1038/sdata.2018.29 (2018).
[22] Nature Scientific Data Editorial, https://www.nature.com/articles/sdata201895

The Identifiers.org resolver provides a stable resolution service for these Compact Identifiers, taking into consideration information such as the uptime and reliability of all available hosting resources. Identifiers.org registry focuses on resources that are of interest, mainly, for the life sciences community. In this field, traditional human-to-data analysis and processing methods gave way to machine-to-data, or proxy, data wrangling techniques and mechanisms. M2M (machine-to-machine) communication is on the rise, and with it, the need for descriptive actionable metadata around objects of processing.

Latest recommendations on metadata mark-up on life sciences resources are driven by Bioschemas[23] profile modelling that is fed back to Schema.org. Best practices on mark-up mechanisms can be found on this document from Google[24], where JSON-LD is the recommended mechanism. Identifiers.org has built a metadata API[25] that allows our users to fetch JSON-LD formatted metadata for both Compact Identifiers and any URL. There are several mechanisms by which data providers can offer JSON-LD metadata, and a very popular one is to embed it in landing pages, e.g. for Compact Identifier 'ensembl:ENSG00000139618', identifiers.org resolution services redirects to this Ensembl landing page[26] (see Figure 18).

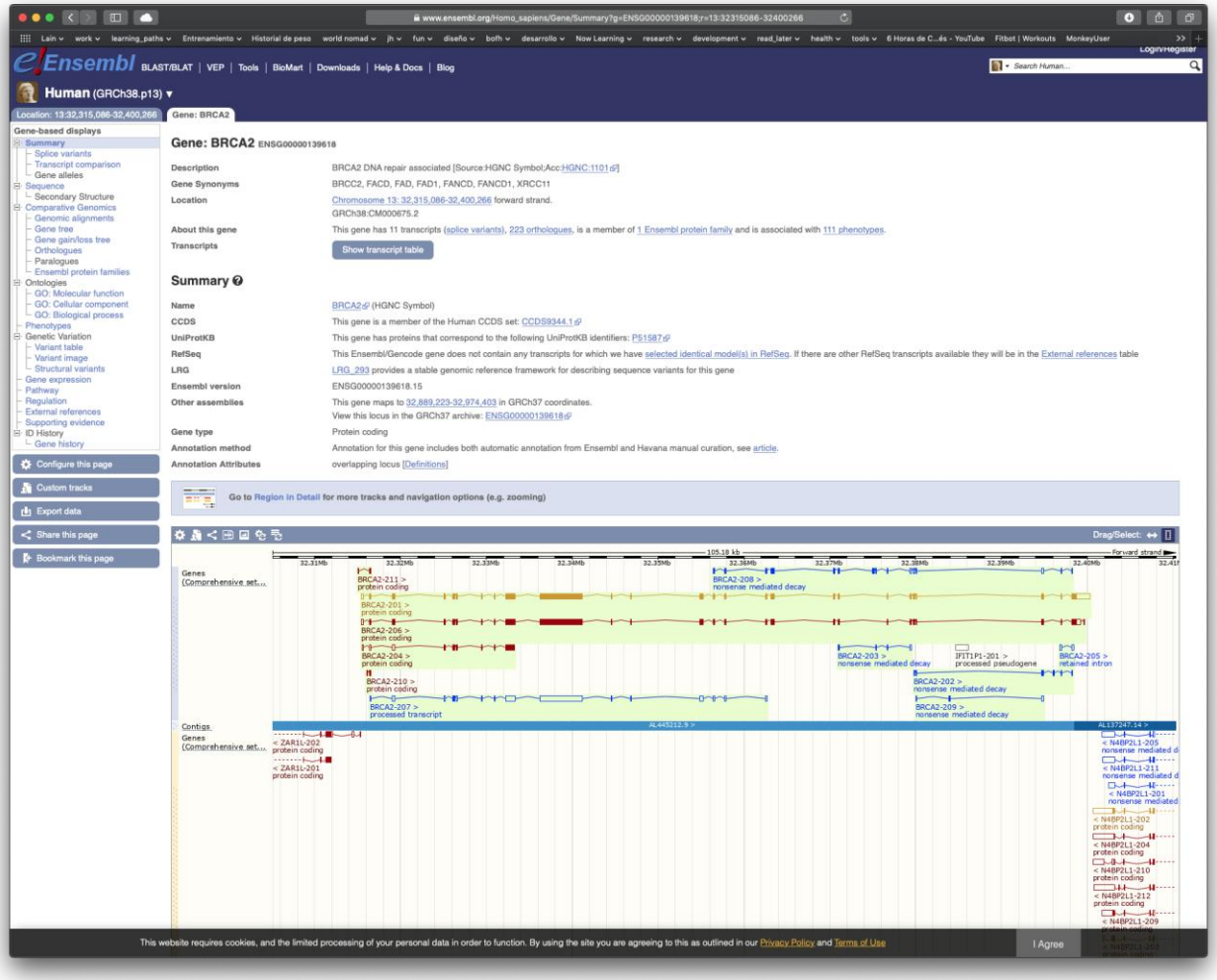


Figure 18 Ensembl landing page for ensembl:ENSG00000139618

[23] https://bioschemas.org/
[24] https://developers.google.com/search/docs/guides/intro-structured-data
[25] https://github.com/identifiers-org/cloud-ws-metadata
[26] https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000139618;r=13:32315086-32400266

By looking at its source code in the browser (see Figure 19), we can see how metadata has been embedded in the landing page in JSON-LD format (see Figure 20).

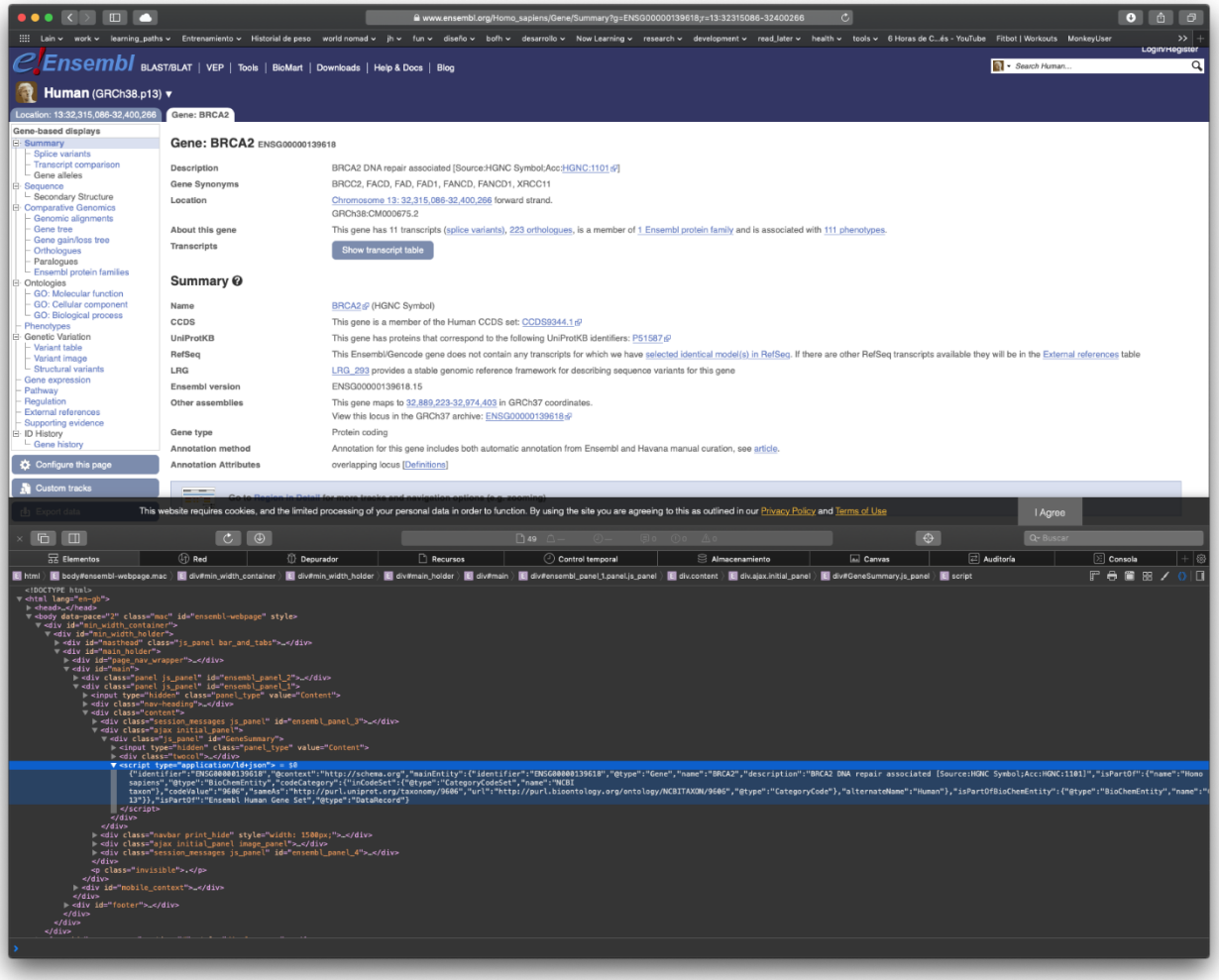
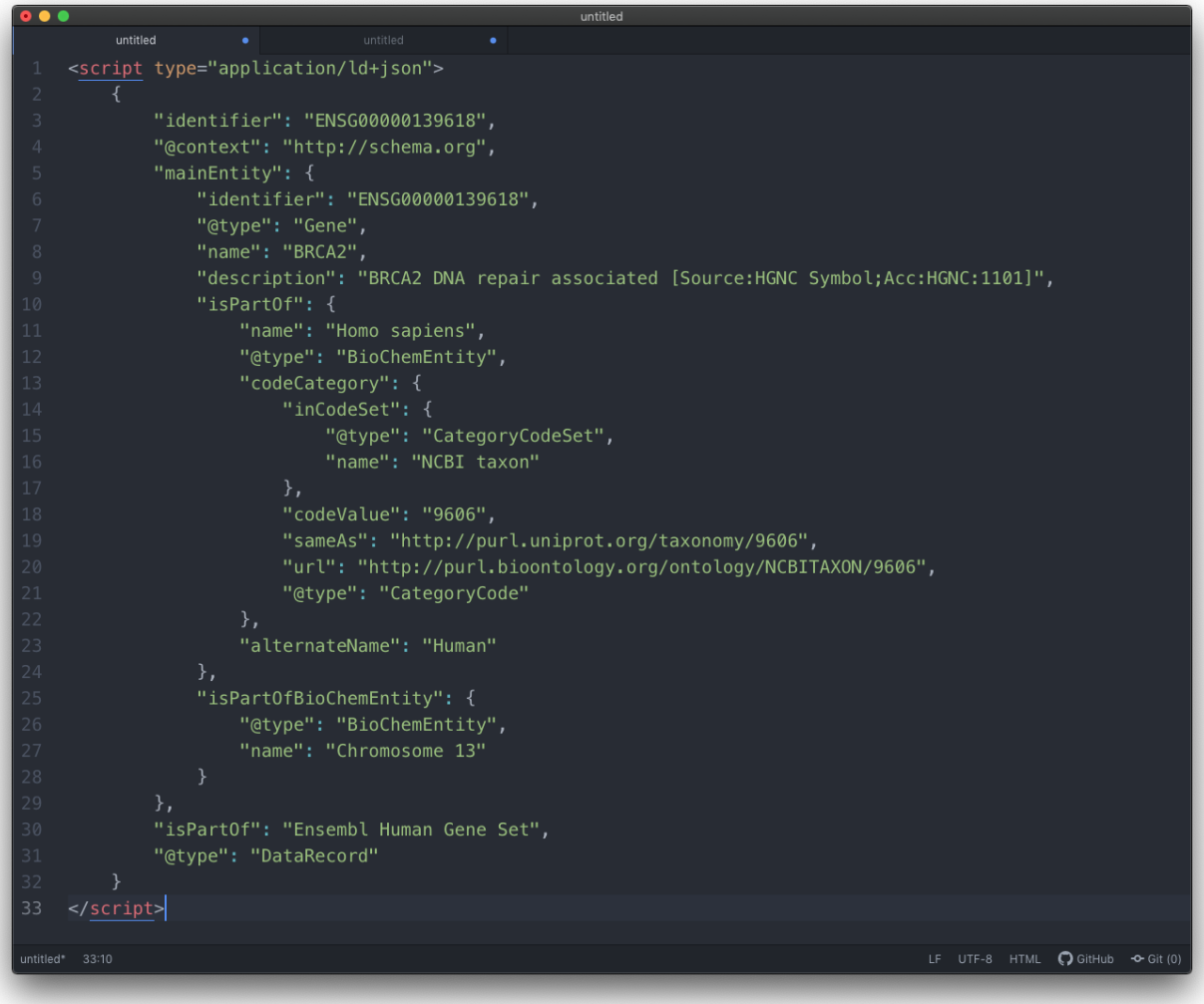


Figure 19 Ensembl landing page source code showing embedded metadata information in JSON-LD format



```
<script type="application/ld+json">
    {
        "identifier": "ENSG00000139618",
        "@context": "http://schema.org",
        "mainEntity": {
            "identifier": "ENSG00000139618",
            "@type": "Gene",
            "name": "BRCA2",
            "description": "BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]",
            "isPartOf": {
                "name": "Homo sapiens",
                "@type": "BioChemEntity",
                "codeCategory": {
                    "inCodeSet": {
                        "@type": "CategoryCodeSet",
                        "name": "NCBI taxon"
                    },
                    "codeValue": "9606",
                    "sameAs": "http://purl.uniprot.org/taxonomy/9606",
                    "url": "http://purl.bioontology.org/ontology/NCBITAXON/9606",
                    "@type": "CategoryCode"
                },
                "alternateName": "Human"
            },
            "isPartOfBioChemEntity": {
                "@type": "BioChemEntity",
                "name": "Chromosome 13"
            }
        },
        "isPartOf": "Ensembl Human Gene Set",
        "@type": "DataRecord"
    }
</script>
```

Figure 20 Details of JSON-LD formatted metadata for Compact Identifier ensembl:ENSG00000139618

Among the different mechanisms that can be used for embedding JSON-LD formatted metadata, most data resources choose to do it dynamically, using JavaScript to inject the information after the page has been loaded. This choice makes metadata extraction very expensive, on both time and space dimensions, as the landing page not only has to be loaded, but also all its associated JavaScript has to be executed, before the extraction begins. Identifiers.org metadata extraction API is available on our production deployment, but in prototype stage because of this time and space complexity, that makes the process of scaling up to attend more requests something we are working on. In the meantime, we offer a method for our community to run the metadata API in the infrastructure of their choice[27] (locally, on-premises, hybrid/multi-cloud environment)

With the increasing adoption of identifiers.org as an in-line data citation mechanism among the life sciences community in mind, we have built a Python Notebook[28] that illustrates how to use identifiers.org metadata API for the purpose of exploring the adoption of JSON-LD formatted metadata among the resources that are currently active in identifiers.org registry.

[27] https://github.com/identifiers-org/cloud-ws-metadata
[28] https://github.com/identifiers-org/metadata-landscape

7.2 Next steps

JSON-LD formatted metadata is the latest recommended annotation mechanism for resources landing pages.

Although the data format itself is lightweight, and very common among internet oriented platforms, the means, chosen by life sciences resources, for implementing this, have an associated computational complexity.

Our next steps, start at evolving our prototype service API to a stage, where the metadata processing and extraction mechanisms can scale in a robust and reliable way, so we can address this challenge.

Identifiers.org's intention is to incorporate this evolved, production ready service API, to its portfolio, and make it available, e.g. via EOSC Portal, for the community to access available metadata associated with compact identifiers, programmatically.

7.3 Lessons learned

The initial findings when working with this API, was the low adoption of metadata annotations by the providers in the registry (see Figure 21), only 5% of the providers.

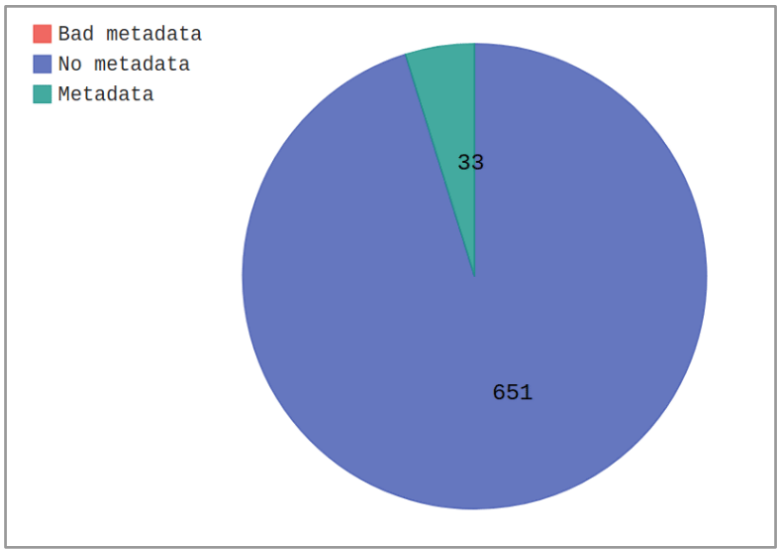


Figure 21 Initial findings on metadata annotation adoption

We decided to cross check these results with Google Structure Data Testing Tool[29].

Our first approach was to run a systematic check on all the resolved URLs for all the explored Compact Identifiers in the registry, but, unfortunately, the tool does not offer an API that can be used for these purposes, so a random subset of Compact Identifiers was selected for manual check between Google's Tool and identifiers.org metadata API[30].

This latter exercise showed that there are more than those 33 providers that offer metadata embedded in their landing pages, e.g. UniProtKB, but through one of the two other mechanisms mentioned in the documentation from Google[31]: RDF and Microdata. In addition, we found that, in some cases, e.g. for

[29] https://search.google.com/structured-data/testing-tool/u/0/?hl=ES
[30] https://github.com/identifiers-org/cloud-ws-metadata
[31] https://developers.google.com/search/docs/guides/intro-structured-data

Compact Identifier ensembl:ENSG00000139618[32], Google Structured Data Testing Tool did not detect any embedded metadata, while our API did report back the metadata content. Clearly, as mentioned in the documentation from Google, while there exists previous metadata annotation mechanisms, i.e. RDF and Microdata, the recommendation is JSON-LD formatted metadata, under the umbrella of Schema.org context definitions, and, although the community is moving towards that implementation (especially for new resources or as an update on those that did not have metadata annotations), it is still a developing aspect in the metadata world.

[32] https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000139618;r=13:32315086-32400266

8 Mechanisms of enhanced provenance information in digital collections

User stories:
As a researcher, I want to be able to understand the content and reuse potential of a dataset, so that I know if and how I can use it in my research.

As a user of EThOS, I want to be able to access and understand information about the funding and institutional fields within the database, so that I can understand how these organisations are represented within the metadata.

In FREYA deliverable D4.2 *Using the PID Graph: Provenance in Disciplinary Systems*[33], provenance was defined for The British Library (BL) as information relating to the origin, source and curation of its collection items both digital and physical. It can also pertain to the source of metadata and documentation about the object.

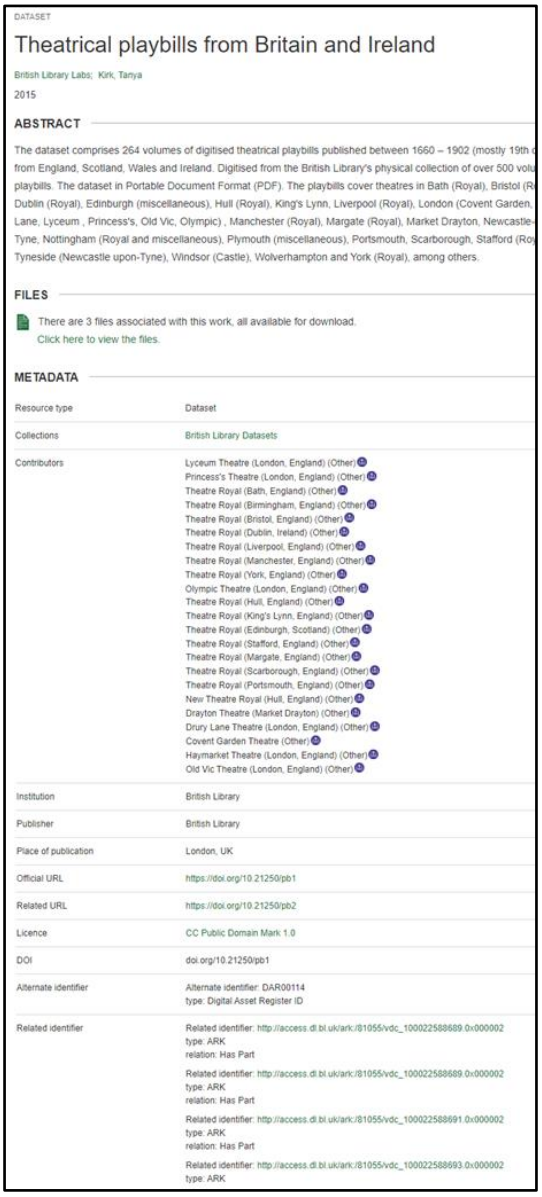
The BL is attempting to enhance the provenance of its existing resources using PIDs in both the resource's repository record and PID metadata and to create workflows for capturing more PIDs for newly added resources, thereby extending the PID Graph for these collections and addressing the BL's provenance-related user stories.

8.1 Solution

8.1.1 Metadata enhancement

For a previous deliverable D4.2 *Using the PID Graph: Provenance in Disciplinary Systems*[34], datasets were augmented with additional provenance metadata. These were selected as they had existing metadata available to be added into the repository record.

[33] https://doi.org/10.5281/zenodo.3249832
[34] https://doi.org/10.5281/zenodo.3249833



DATASET

Theatrical playbills from Britain and Ireland

British Library Labs; Kirk, Tanya

2015

ABSTRACT

The dataset comprises 264 volumes of digitised theatrical playbills published between 1660 – 1902 (mostly 19th c
from England, Scotland, Wales and Ireland. Digitised from the British Library's physical collection of over 500 volu
playbills. The dataset in Portable Document Format (PDF). The playbills cover theatres in Bath (Royal), Bristol (R
Dublin (Royal), Edinburgh (miscellaneous), Hull (Royal), King's Lynn, Liverpool (Royal), London (Covent Garden,
Lane, Lyceum , Princess's, Old Vic, Olympic) , Manchester (Royal), Margate (Royal), Market Drayton, Newcastle-
Tyne, Nottingham (Royal and miscellaneous), Plymouth (miscellaneous), Portsmouth, Scarborough, Stafford (Roy
Tyneside (Newcastle upon-Tyne), Windsor (Castle), Wolverhampton and York (Royal), among others.

FILES

There are 3 files associated with this work, all available for download.
Click here to view the files.

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)
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
DOI	doi.org/10.21250/pb1
Alternate identifier	Alternate identifier: DAR00114 type: Digital Asset Register ID
Related identifier	Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588689.0x000002 type: ARK relation: Has Part Related identifier: http://access.dl.bl.uk/ark:/81055/vdc_100022588689.0x000002 type: ARK relation: Has Part 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_100022588693.0x000002 type: ARK

Figure 22 A record augmented with identifiers

As an additional step, in December 2019, we added new datasets to data.bl.uk which were derived from an existing dataset, *Digitised 19th Century Books - Metadata - 01/09/2013*[35] and created DataCite DOIs for them. These incorporated the relevant related identifier in their metadata as well as supporting information about the derivation methodology in the metadata where available.

[35] https://doi.org/10.21250/DB21

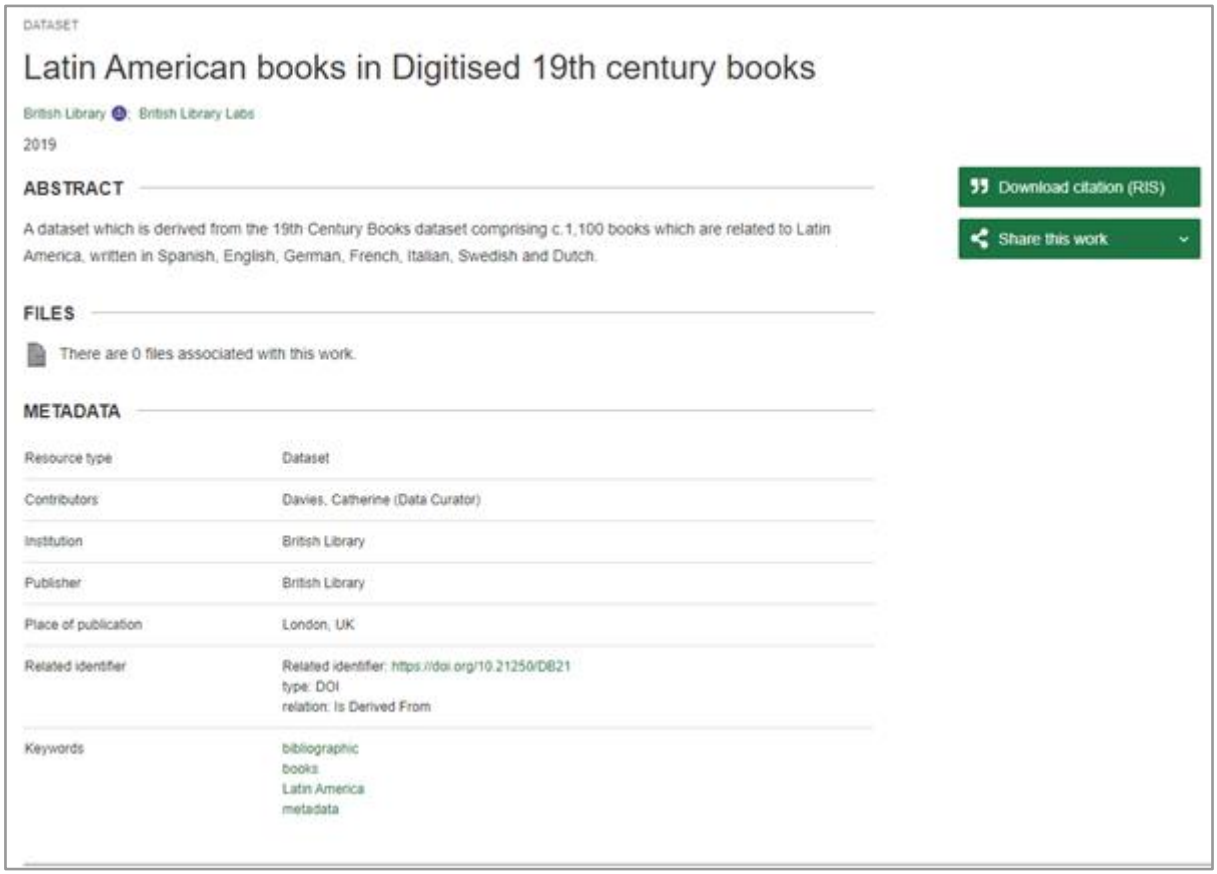
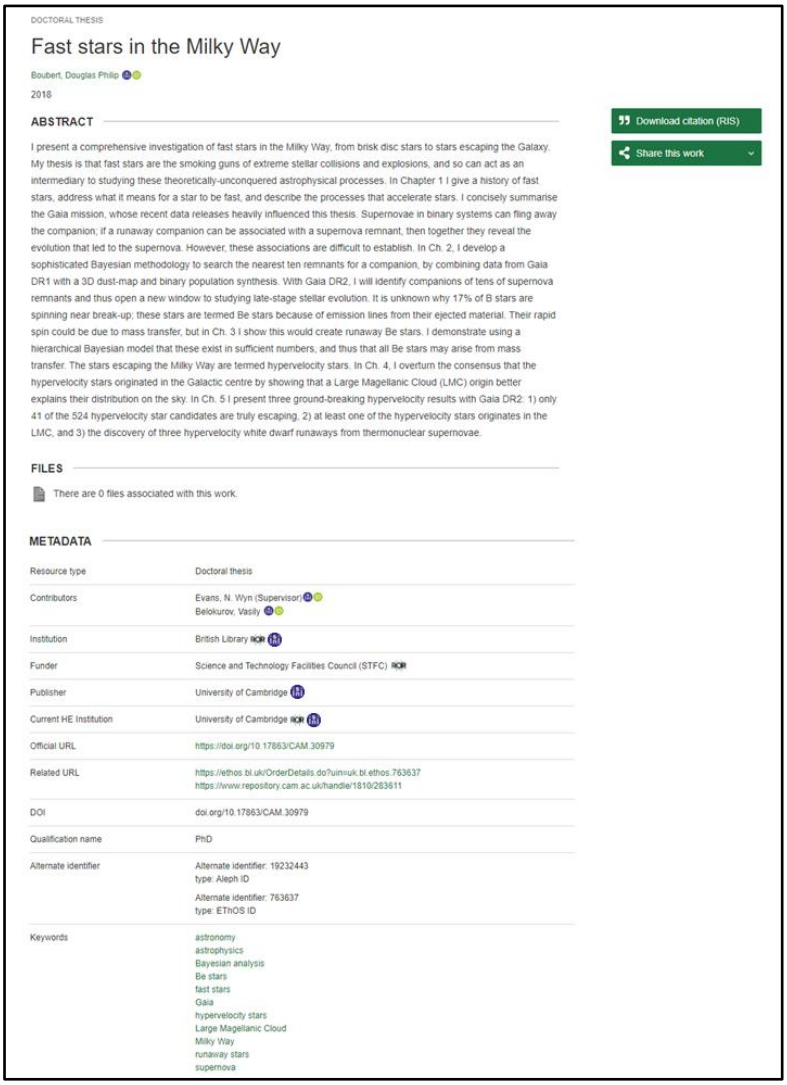


Figure 23 Dataset record with provenance information related to its "parent dataset".

For the EThOS collection, the index of UK Doctoral theses, the BL is working to augment the metadata ahead of its migration to a new platform in 2020/21. Some example records were added to the Demo repository and various preparatory actions have been taken on the metadata of the whole collection, the results of which are available via the British Library's repository.[36]

[36] https://doi.org/10.23636/1156



DOCTORAL THESIS

Fast stars in the Milky Way

Boubert, Douglas Philip

2018

Download citation (RIS)

Share this work

ABSTRACT

I present a comprehensive investigation of fast stars in the Milky Way, from brisk disc stars to stars escaping the Galaxy. My thesis is that fast stars are the smoking guns of extreme stellar collisions and explosions, and so can act as an intermediary to studying these theoretically-unconquered astrophysical processes. In Chapter 1 I give a history of fast stars, address what it means for a star to be fast, and describe the processes that accelerate stars. I concisely summarise the Gaia mission, whose recent data releases heavily influenced this thesis. Supernovae in binary systems can fling away the companion; if a runaway companion can be associated with a supernova remnant, then together they reveal the evolution that led to the supernova. However, these associations are difficult to establish. In Ch. 2, I develop a sophisticated Bayesian methodology to search the nearest ten remnants for a companion, by combining data from Gaia DR1 with a 3D dust-map and binary population synthesis. With Gaia DR2, I will identify companions of tens of supernova remnants and thus open a new window to studying late-stage stellar evolution. It is unknown why 17% of B stars are spinning near break-up; these stars are termed Be stars because of emission lines from their ejected material. Their rapid spin could be due to mass transfer, but in Ch. 3 I show this would create runaway Be stars. I demonstrate using a hierarchical Bayesian model that these exist in sufficient numbers, and thus that all Be stars may arise from mass transfer. The stars escaping the Milky Way are termed hypervelocity stars. In Ch. 4, I overturn the consensus that the hypervelocity stars originated in the Galactic centre by showing that a Large Magellanic Cloud (LMC) origin better explains their distribution on the sky. In Ch. 5 I present three ground-breaking hypervelocity results with Gaia DR2: 1) only 41 of the 524 hypervelocity star candidates are truly escaping, 2) at least one of the hypervelocity stars originates in the LMC, and 3) the discovery of three hypervelocity white dwarf runaways from thermonuclear supernovae.

FILES

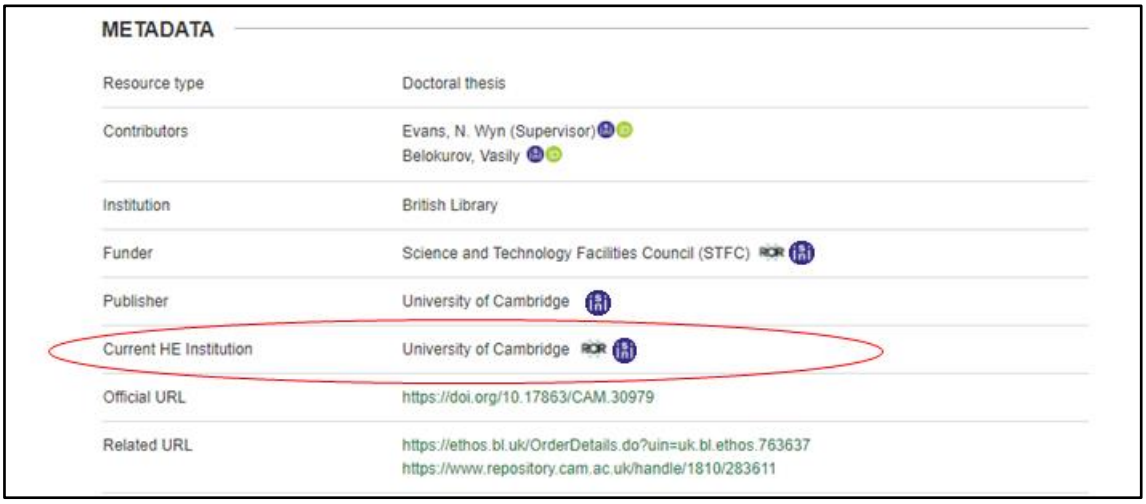
There are 0 files associated with this work.

METADATA

Resource type	Doctoral thesis
Contributors	Evans, N. Wyn (Supervisor) Belokurov, Vasily
Institution	British Library
Funder	Science and Technology Facilities Council (STFC)
Publisher	University of Cambridge
Current HE Institution	University of Cambridge
Official URL	https://doi.org/10.17863/CAM.30979
Related URL	https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.763637 https://www.repository.cam.ac.uk/handle/1810/283611
DOI	doi.org/10.17863/CAM.30979
Qualification name	PhD
Alternate identifier	Alternate identifier: 19232443 type: Aleph ID Alternate identifier: 763637 type: EThOS ID
Keywords	astronomy astrophysics Bayesian analysis Be stars fast stars Gaia hypervelocity stars Large Magellanic Cloud Milky Way runaway stars supernova

Figure 24 A mock-up of a record in the new EThOS platform

Fields with multiple values due to the current platform's limitations including Funder and Supervisor were separated in preparation for migration and the following preparatory migration steps were taken.



METADATA

Resource type	Doctoral thesis
Contributors	Evans, N. Wyn (Supervisor) Belokurov, Vasily
Institution	British Library
Funder	Science and Technology Facilities Council (STFC)
Publisher	University of Cambridge
Current HE Institution	University of Cambridge
Official URL	https://doi.org/10.17863/CAM.30979
Related URL	https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.763637 https://www.repository.cam.ac.uk/handle/1810/283611

Figure 25 A mock up of the metadata for an EThOS record highlighting the Current HE Institution field which will be able to support organisational identifiers such as ROR and ISNI

The Current HE Institution field was matched with ROR and ISNI. The success rate of the matching varied across the different fields. For the current institution, as this is already a controlled list in EThOS, this had the highest success rate. 141 of 143 institutions all matched with ISNI, and ISNIs were created for the remaining institutions by the ISNI team at the BL. Eight did not have ROR IDs but these were where the BL regards Institutes of the University of London as individual institutions, which ROR does not.

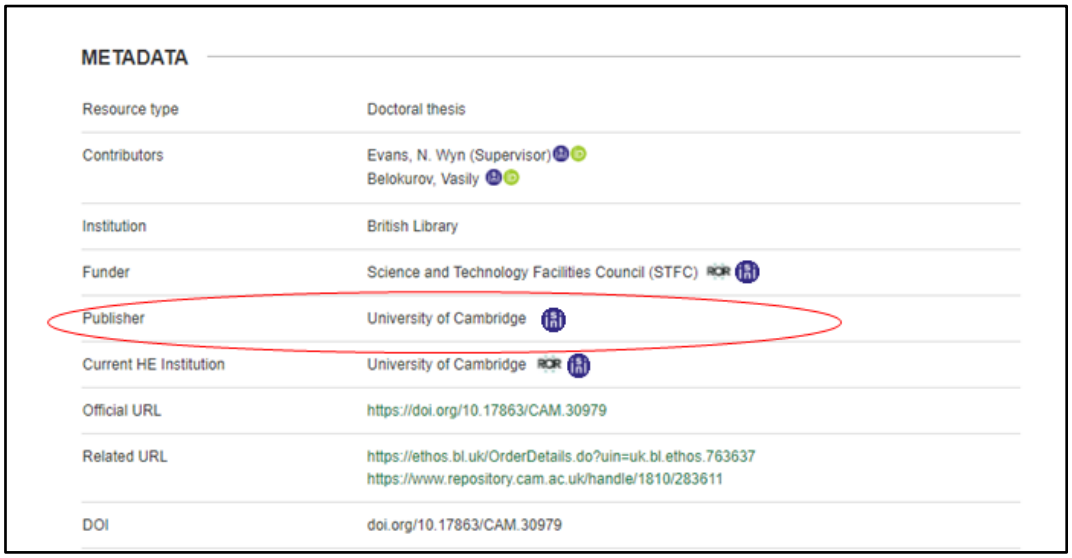


Figure 26 A mock up of the metadata for an EThOS record highlighting the Publisher field which will be able to support organisational identifiers such as ISNI

Awarding Institution, here called Publisher, was matched with ROR. The Awarding Institution or Publisher field is a free text field in EThOS at present. OpenRefine and the ROR reconciler were used to match against the ROR database, and the percentage of matches was 98%. Because the Awarding Institution does not change over time, as the current institution does, in the event of mergers and/or closures of higher education institutions, a lower match rate was expected. ROR does not hold historical information about institutions such as former names, therefore ISNI has always been considered a more suitable use case for this field. However, given the high match rather this may be reconsidered. Matching against ISNI was not undertaken due to resource constraints across the ISNI team, but it can be assumed that matching would be similarly high.

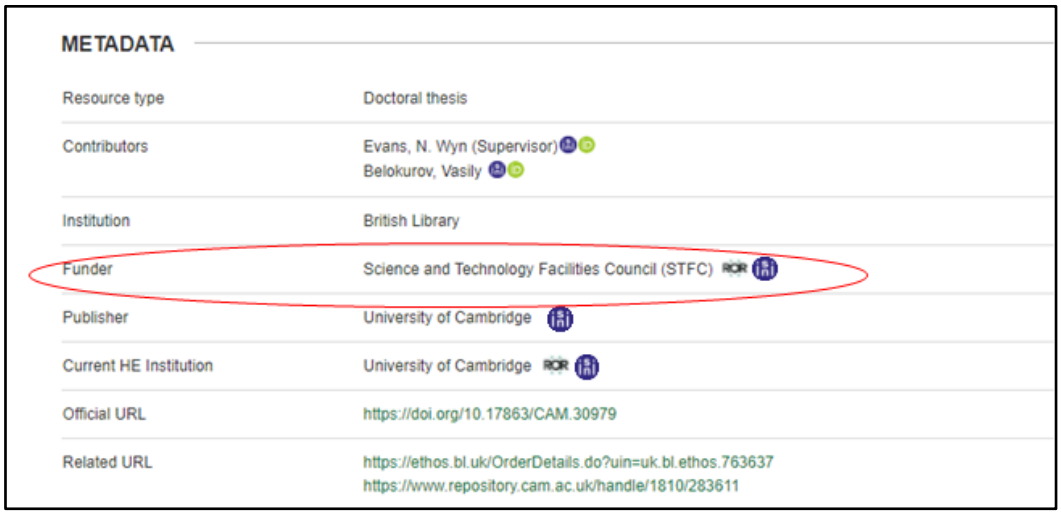


Figure 27 A mock-up of the metadata for an EThOS record highlighting the Funder field which will be able to support organisational identifiers such as ROR

The Funder field was matched with ROR. The Funder field is populated for 6% of records. The plan for the repository is to introduce a Crossref Funder Registry look-up. However, as an experiment in attempting to gauge the cleanliness of the Funder data in EThOS, as it is again a free text field, the matching had a 37% success rate and was a very manual process due to variations within the fields.

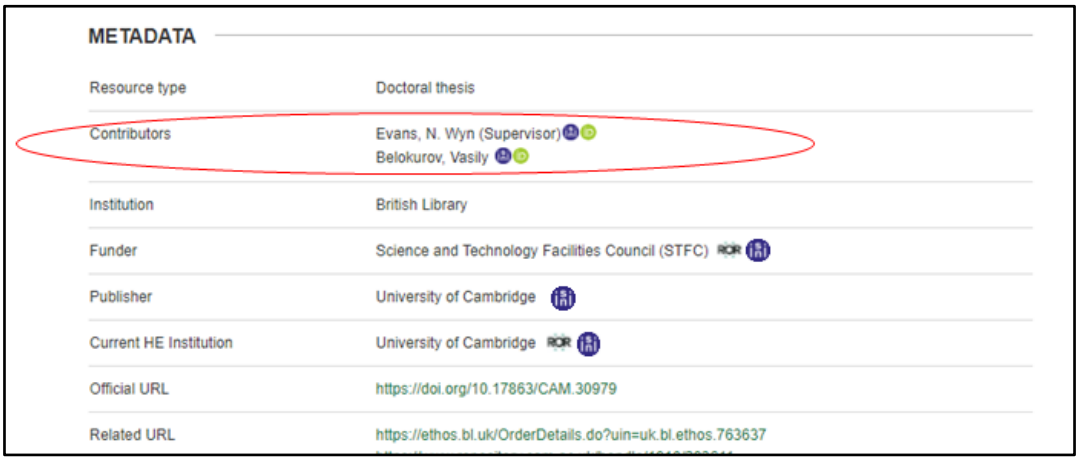


Figure 28 A mock-up of the metadata for an EThOS record highlighting the Contributor field which will be able to support identifiers such as ISNI and ORCID

The current EThOS platform supports supervisors in one single field, which are included in approximately 20% of records, but identifiers such as ORCID and ISNI are not supported for them. This field was split to identify individual names which could be matched against the ISNI and ORCID registries. The ISNI Quality Team at the BL attempted to match a sample of 100 supervisors against the database, however, as it was difficult to draw a conclusive relationship between the work for which there was metadata and the supervisor, it was not possible to make a conclusive match.

8.1.2 Workflow development

The development of formal workflows which enable the capturing of provenance metadata has proved somewhat challenging due to the extremely varied nature of datasets within the data.bl.uk collection. It is expected that by leading by example, new records added to the repository will contain rich provenance metadata. A selection of derived datasets has now been added to demonstrate this.[37] However, in several cases within the data.bl.uk collection this metadata can be hard to find or does not have a suitable identifier for inclusion in the metadata.

8.2 Next steps

All of these additions to the metadata have highlighted the need for good UX design of the repository including the capability to manage large numbers of identifiers, as well as accommodating a variety of identifiers. These features were described in FREYA deliverables 4.2 and 4.4, and will be delivered in 2020.

As the Shared Research Repository is developed, workflows are being established for how items are added to it. As part of that work, increased awareness of identifiers is required in order to utilise this new functionality which can be incorporated into the new EThOS platform once it is migrated. This will be developed through using a Crossref Funder Registry look-up and developing controlled vocabularies wherever possible.

It is still undecided how these identifiers will be displayed, but this will be worked out in scoping by the development partner. There is a commitment that this information will be included in any DOI metadata created from data.bl.uk datasets. In order to improve the representation of supervisor's identifiers in the

[37] https://bl.iro.bl.uk/collection/36116aa1-7037-40f3-9b91-ecb1be15e226

metadata, a fresh attempt at matching supervisors with the ISNI database utilising subject headings as a cross reference will be attempted.

8.3 Lessons learned

One of the issues with this work was that it made the records very long and possibly unusable. In an earlier version of the user interface the files were only available for download at the bottom of the screen. We are planning to improve the display in records to accommodate the larger number of identifiers.

Figure 29 The full record of Theatrical Playbills from Great Britain and Ireland. The length is due to the number of related identifiers which are cited in the record.

Due to the early stages of this implementation work, further lessons learned are somewhat limited but can be provided at a later stage in the process.

9 Conclusion

FREYA's New PID Types work package (WP3) set out to explore new PID types and services that would be useful additions to the PID Graph. Throughout the course of the work package, FREYA partners have surveyed the landscape of existing and needed PIDs (D3.1), gathered user stories and requirements from the wider PID community (D3.2), and now carried out prototype implementations of those PIDs that were deemed suitable candidates (described in this deliverable). As this exploratory work package concludes, the results and resulting prototype implementations will be taken up by the work package responsible for integrating the PID Graph (WP4), which will leverage the disciplinary expertise of the FREYA partners to build on these more foundational explorations in order to more robustly populate their respective areas of the PID Graph.

FREYA partners have explored the creation of four new PID types: PIDs for scientific instruments, PIDS for scientific facilities, PIDs for organisations and grant IDs. PIDs for organisations is the most complete from the FREYA perspective, with the ROR registry established and the ROR ID incorporated into DataCite services, but there are still questions outside the scope of the FREYA project, for instance concerning the sustainability of ROR. Grant IDs and PIDs for scientific instruments both made good progress during the FREYA project so far, with demonstrable outputs showcasing their possibilities from a user's perspective. Work will continue beyond FREYA to coordinate with other partners and entities external to the project, so that identified bottleneck issues, such as incorporating instrument PIDs into the DataCite Metadata Schema or encouraging funders to register DOIs for their grants, can be resolved. PIDs for scientific facilities are still under consideration, though early efforts proved challenging in part due to the multifaceted nature of facilities as funders, instruments and organisations, PIDs for which are largely being addressed elsewhere. Work in this area will continue as part of FREYA WP4.

In addition, FREYA partners have investigated new services for existing PIDs, focusing primarily on DOI registration workflows and metadata improvement. PANGAEA and Identifiers.org both explored ways to improve initial PID registration without significant human intervention by making use of Schema.org and JSON-LD embedded in landing pages. In both cases, this work was successful. The British Library explored enhancing their digital collections with additional provenance metadata, as well as the workflows to achieve this. This work was also successful, though future work will be conducted to explore how to make the newly detailed records seem less unwieldy to the end user.

Along the way, it has become clear that developing and implementing new PID types and new PID services is very much a community effort, requiring significant coordination between multiple players to escort a nascent PID from idea to broader uptake. First, a new PID must be designed, which is itself a significant undertaking, involving decisions about resolution, metadata hosting, centralised infrastructure, sustainable costs and so forth, not to mention user research about which problems the PID is meant to solve and how it does or does not address the needs of multiple communities. Assuming this PID design process is successful, the PID must also be absorbed into various metadata schemas, which may be on their own update schedules independent of introducing new PIDs. For many creators of PIDs, an API that follows the newly updated metadata schema is not sufficient for their needs, so user interfaces must be designed to aid in PID creation. And finally, the results of all of these new PID types and more detailed PID connections must be available to all end users, requiring additional user interface updates and tweaks. Projects like FREYA allow for the close collaboration of disparate entities involved at various stages of the PID implementation pipeline in order to see the completion of prototypes and to provide a context in which experimentation, with its possibility of prototype abandonment, can be supported. In this case, we are pleased to report that though some of these prototypes surfaced challenges that will not be resolved within the course of FREYA, none are to be wholly abandoned. The shape they may take as they are integrated into the PID Graph may evolve over the remaining month of FREYA and beyond the project timeline, but all have made valuable inroads toward our shared vision of a world of interconnected PIDs.