IJDC | *General Article*

DataCite: Lessons Learned on Persistent Identifiers for Research Data

Laura Rueda
DataCite

Martin Fenner
DataCite

Patricia Cruse
DataCite

Abstract

Data are the infrastructure of science and they serve as the groundwork for scientific pursuits. Data publication has emerged as a game-changing breakthrough in scholarly communication. Data form the outputs of research but also are a gateway to new hypotheses, enabling new scientific insights and driving innovation. And yet stakeholders across the scholarly ecosystem, including practitioners, institutions, and funders of scientific research are increasingly concerned about the lack of sharing and reuse of research data. Across disciplines and countries, researchers, funders, and publishers are pushing for a more effective research environment, minimizing the duplication of work and maximizing the interaction between researchers. Availability, discoverability, and reproducibility of research outputs are key factors to support data reuse and make possible this new environment of highly collaborative research.

An interoperable e-infrastructure is imperative in order to develop new platforms and services for to data publication and reuse. DataCite has been working to establish and promote methods to locate, identify and share information about research data. Along with service development, DataCite supports and advocates for the standards behind persistent identifiers (in particular DOIs, Digital Object Identifiers) for data and other research outputs. Persistent identifiers allow different platforms to exchange information consistently and unambiguously and provide a reliable way to track citations and reuse. Because of this, data publication can become a reality from a technical standpoint, but the adoption of data publication and data citation as a practice by researchers is still in its early stages.

Since 2009, DataCite has been developing a series of tools and services to foster the adoption of data publication and citation among the research community. Through the years, DataCite has worked in a close collaboration with interdisciplinary partners on these issues and we have gained insight into the development of data publication workflows. This paper describes the types of different actions and the lessons learned by DataCite.

Received 20 October 2015 ~ *Revised* 28 June 2016 ~ *Accepted* 28 June 2016

Correspondence should be addressed to Laura Rueda, Welfengarten 1B, 30167 Hannover (Germany). Email: laura.rueda@datacite.org

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

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

 doi:10.2218/ijdc.v11i2.421

Introduction

Data are the infrastructure of science and they serve as the groundwork for scientific pursuits. Data publication has emerged as a game-changing breakthrough in scholarly communication. Data form the outputs of research but also are a gateway to new hypotheses, enabling new scientific insights and driving innovation. And yet stakeholders across the scholarly ecosystem, including practitioners, institutions and funders of scientific research, are increasingly concerned about the lack of sharing and reuse of research data. Across disciplines and countries, researchers, funders, and publishers are pushing for a more effective research environment, minimizing the duplication of work and maximizing the interaction between researchers. Availability, discoverability and reproducibility of research outputs are key factors to support data reuse and make possible this new environment of highly collaborative research.

An interoperable e-infrastructure is imperative in order to develop new platforms and services for data publication and reuse. DataCite has been working to establish and promote methods to locate, identify and share information about research data. Along with service development, DataCite supports and advocates for the standards behind persistent identifiers (in particular DOIs, Digital Object Identifiers) for data and other research outputs. Persistent identifiers allow different platforms to exchange information consistently and unambiguously and provide a reliable way to track citations and reuse. Because of this, data publication can become a reality from a technical standpoint, but the adoption of data publication and data citation as a practice by researchers is still in its early stages.

As the broader stakeholder community works to make research data as essential as traditional publications, it is crucial to develop services to meet the researchers' needs and to provide them with services to understand the impact of all their research outputs, including data.

Since 2009, DataCite has been developing a series of tools and services to foster the adoption of data publication and citation among the research community. Through the years, DataCite has worked in a close collaboration with interdisciplinary partners on these issues and we have gained insight into the development of data publication workflows. This paper describes the types of different actions and the lessons learned by DataCite.

Integration Workflows

DataCite works collaboratively with our members[1] to deliver services to more than 600 data centres. Each data centre has developed different infrastructures and workflows that are designed to best meet their community's needs. Some are small and others serve thousands of users. Independent of the platform they use (e.g. a well-known repository software, an ad-hoc solution or a mixed approach), our experience has taught us to encourage the early involvement of three stakeholder communities: IT experts and developers, librarians and information scientists, and user experience designers (see Figure 1). Each of these stakeholders brings a unique perspective that is crucial in

1 DataCite Members: https://www.datacite.org/about-datacite/members

delivering a successful service: a reliable technical infrastructure, high-quality metadata and curation, and an adequate solid interface to interact with the platform.

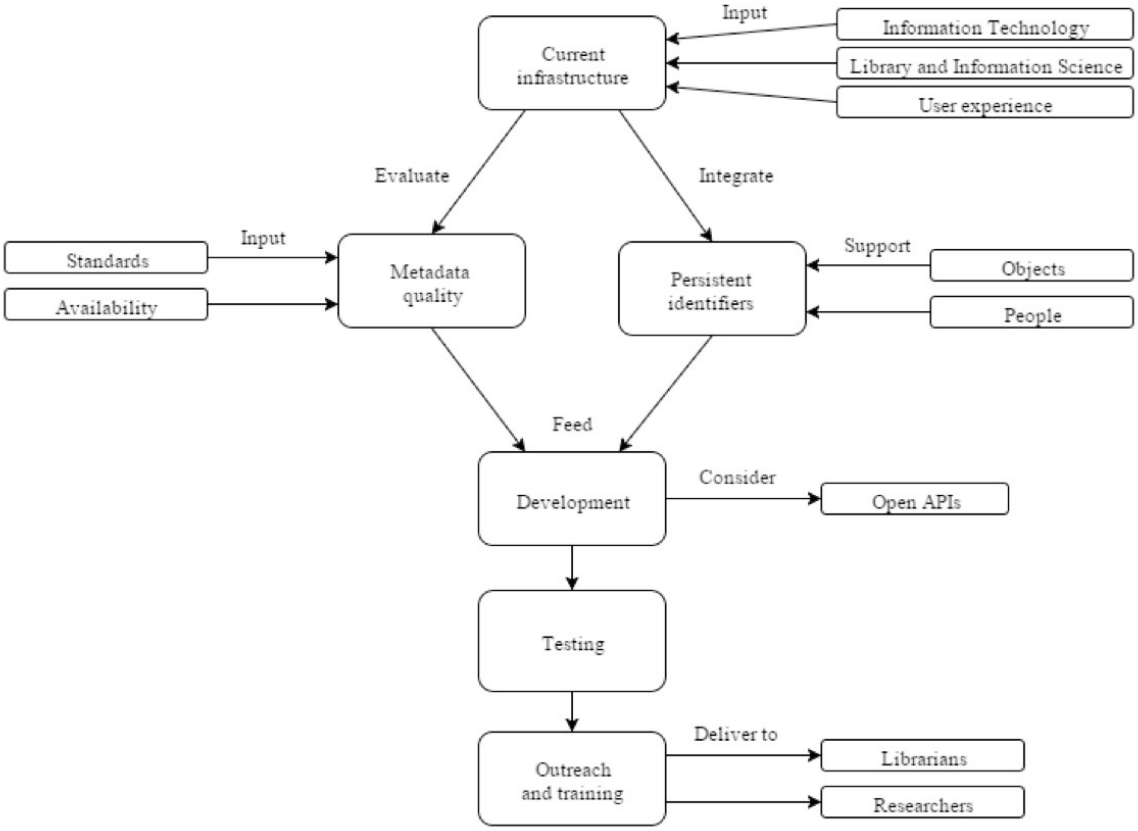

Figure 1. Development and integration workflow.

Excluding any of the stakeholders' perspectives can be problematic and create problems that are difficult to solve retrospectively. Providing persistent identifiers for data implies a set of commitments. Unstable or not scalable services are usually rebuilt or redesigned shortly after they prove themselves to be unreliable and thereby demanding extra effort to keep persistent identifiers linked resolvable to content. Low-quality metadata, uncurated content, and a lack of internal and/or external organisation create repositories that are impossible to navigate or to obtain information from. This problem directly relates to cumbersome or inadequate interfaces and ultimately discourages researchers from using the service.

To better understand the gaps in a service a good place to begin is an assessment of the existing infrastructure:

1. Consider if the metadata is robust enough to provide an accurate description of the objects,
2. Evaluate if the metadata is compatible with the DataCite's Metadata Schema,[2]
3. Evaluate the persistence of the links or the mechanism to trigger updates, etc.

2 DataCite Metadata Schema: http://schema.datacite.org

Once the assessment is complete and the service is ready and begins assigning DOIs to datasets, interoperability opportunities emerge. However, an open API is imperative so other services can build or exchange information. With a goal of supporting smaller data centres, DataCite maintains an OAI-PMH[3] interface for all its content and also develops a centralized search portal described in the next section.

When all the pieces are in place and the entire infrastructure has been tested then it is important to engage in outreach and training activities. There are multiple ways to encourage the adoption of data publication (i.e. carrot and stick approach), but a sound infrastructure that connects data to publications and to authors is an important piece in publishing data.

Information Exchange Between Services

Persistent identifiers allow independent platforms to interoperate and exchange information. Propagating information results in a complete service and simplifies workflows. Different data repositories and publishers have started adding the persistent identifiers of related content to its own metadata, completing a graph to navigate the entire research landscape (see Figure 2).

A prominent example is the collaboration between Crossref[4], DataCite and ORCID[5] to automatically update ORCID records when a new research output gets assigned a DOI. This integration provides an easy way to keep all the main services up to date and helps propagate the metadata further. In this way, researchers can also keep their publication lists up to date with minimal effort.

In a space where multiple persistent identifier systems seem to compete for adoption, DataCite believes cooperation provides a solid path for success. Two different EU-funded projects (ODIN, the ORCID and DataCite Interoperability Network[6]; and THOR, the Technical and Human Infrastructure for Open Research[7]), have worked to design a thin layer of interoperability (Vision, Kaye, and Aryani, 2014), making independent standards compatible and collectively have developed strategies to unify workflows across different communities. DataCite's Metadata Schema has been compared with multiple standards (ORCID, Dublin Core, CASRAI, MODS, DDI…) regarding contributors, organizations, and objects, to map the pain points and propose modifications (Fenner et al., 2015). THOR's success in this area is because we focused on evaluating how to align our efforts before we move towards each integration.

3 Open Archives Initiative – Protocol for Metadata Harvesting: https://www.openarchives.org/pmh
4 CrossRef: http://crossref.org
5 ORCID: http://orcid.org
6 ODIN Project: http://odin-project.eu
7 THOR Project: http://project-thor.eu

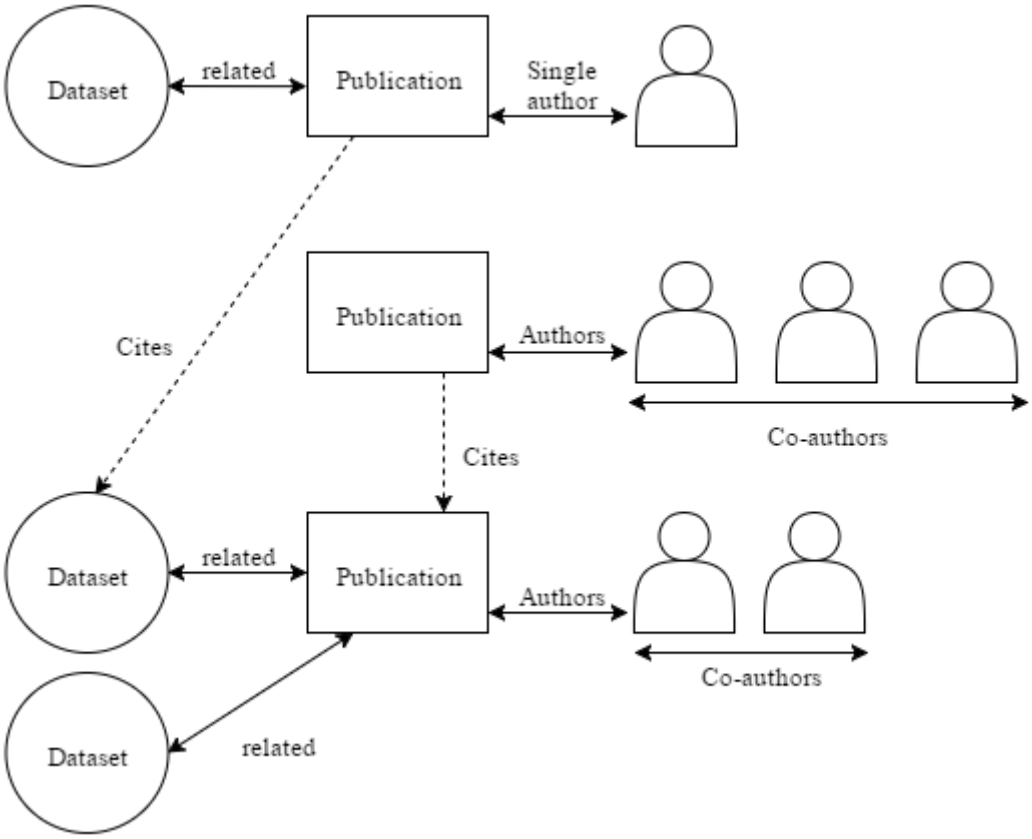

Figure 2. Graph of authors, publications, datasets and their relationships.

Complete and Interoperable Metadata

Quality metadata is imperative to enable researchers to search and find data. DataCite's Metadata Schema defines a mandatory subset of metadata, but more importantly, it provides a wide range of attributes carefully designed, iterated and improved metadata schema for compatibility across systems and standards (Table 1).

As previously mentioned, the DataCite metadata schema has been stress-tested by multiple studies and through practical integrations. Still, we can identify gaps in particular areas, for example how to represent organizations, funding information and contributor roles. The THOR Project is currently working to propose the best approach to improve those properties.

All the metadata stored by DataCite is freely accessible and allows third-party services to build both community-specific and general portals to enhance provided metadata. DataCite cultivates a culture of openness and encourages those integrations as the most effective way to develop data publication and citation practices.

DataCite is also providing a centralised metadata search service, showcasing the flexibility of the system. The quality of the search results relies on the metadata provided by each data centre, and the quantity and quality of metadata can vary dramatically from one to record to another. Different situations limit the metadata provided, including development capacity, licenses, community practices, but it is essential to encourage all data centres to provide as much metadata as possible.

 doi:10.2218/ijdc.v11i2.421

Table 1. DataCite Metadata Schema properties.

Mandatory	Recommended	Optional
Identifier	Subject	Language
Creator	Contributor	Alternate Identifier
Title	Date	Size
Publisher	Related Identifier	Format
Publication Year	Description	Version
Resource Type	Geolocation	Rights

The effort to produce complete and interoperable metadata in addition to persistent identifiers is imperative to make research data as essential as traditional publications. It enables the most basic services: display, exchange, search and reuse.

Data-Level Metrics

The impact of research output can be measured in different ways, using formal citations, altmetrics, and other practices. When it comes to data, it is important to help researchers understand the impact of their work in different contexts. The project Making Data Count (Kratz and Strasser, 2015) developed a tool to collect the online activity surrounding datasets from usage to references, social shares, discussions, and citations.

Currently, researchers are not consistently citing datasets in journal articles, but when they do there is a great degree of variance in their practice. Datasets are mentioned in different places within a journal article: as part of the main body of the journal article, text embedded in the journal article, in the notes section, as a footnote, or within a dedicated section of the journal article, but infrequently in the cited in the references section of a journal article.

To overcome this challenge DataCite is part of an emerging effort to establish a standard practice that would enable a propagation of research data citation. DataCite is also hosting the Data-Level Metrics service from the Making Data Count project and is integrating metrics with other partners. The service is able to perform a full-text search across the journal content, looking for any mention of a dataset via a persistent identifier regardless of its location. Depending on the availability of the information (using different APIs) it also counts views, downloads, and discussions in social media. The DataCite Search service is also using this information to enrich its results (see Figure 3).

Research impact and performance are often evaluated through metrics, principally through citations. As multiple funding agencies and governments have started encouraging and increasingly requiring the publication of datasets, the development of these tools to measure impact can provide incentives for researchers to follow best practices to better understand the impact of their research.

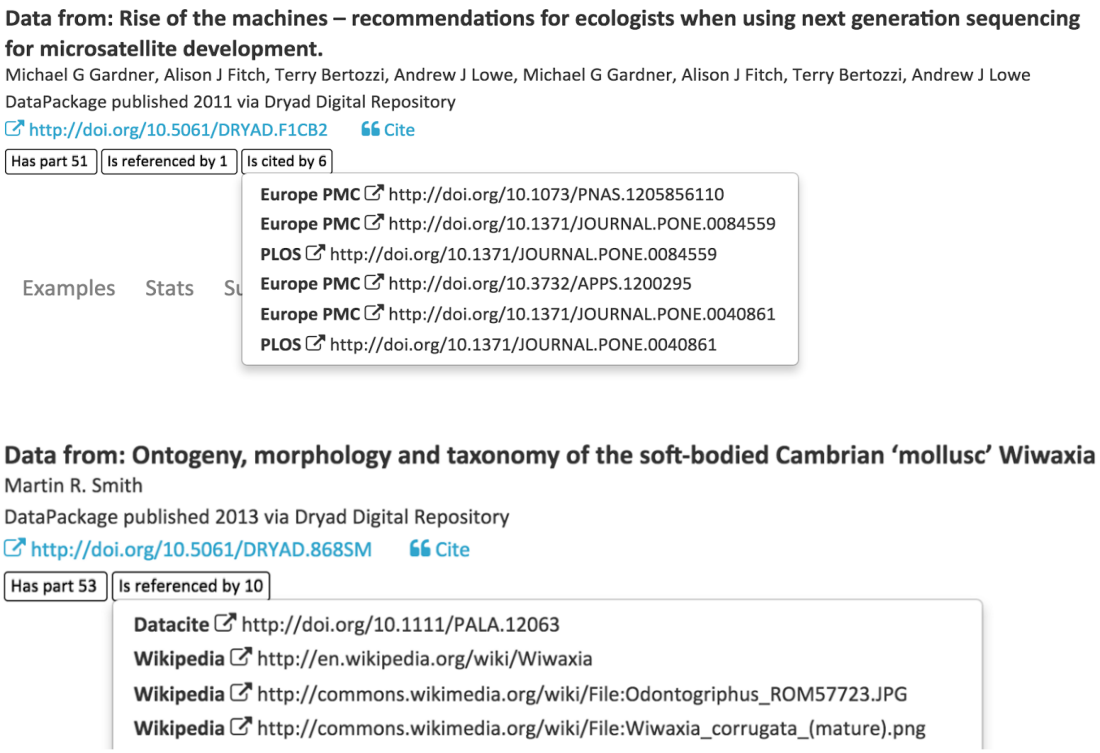

Figure 3. Search results using information from the Data-Level Metrics Service.

Communication for Adoption

With all these tools in place, DataCite is working to support data publication, search, citation and impact tracking. The final step of the process is to provide a scalable plan to reach different communities and engage them to adopt these new services.

As part of the THOR project, DataCite is working collaboratively on the adoption of data as part of research workflows. Through a comprehensive list of training and outreach actions, including a community of ambassadors, our efforts will help showcase data as an important part of every researcher's output.

The future of data publication demands connected services, reliable tools and learning pathways. Through the development of these services, DataCite has learned the importance of interaction with the community to design tools to meet the real needs of our stakeholder community.

The design of our communication engagement strategy includes categorizing our activities based on the stakeholders we wanted to reach. Each stakeholder has different needs and requires specific information to meet those needs (Figure 4).

 doi:10.2218/ijdc.v11i2.421

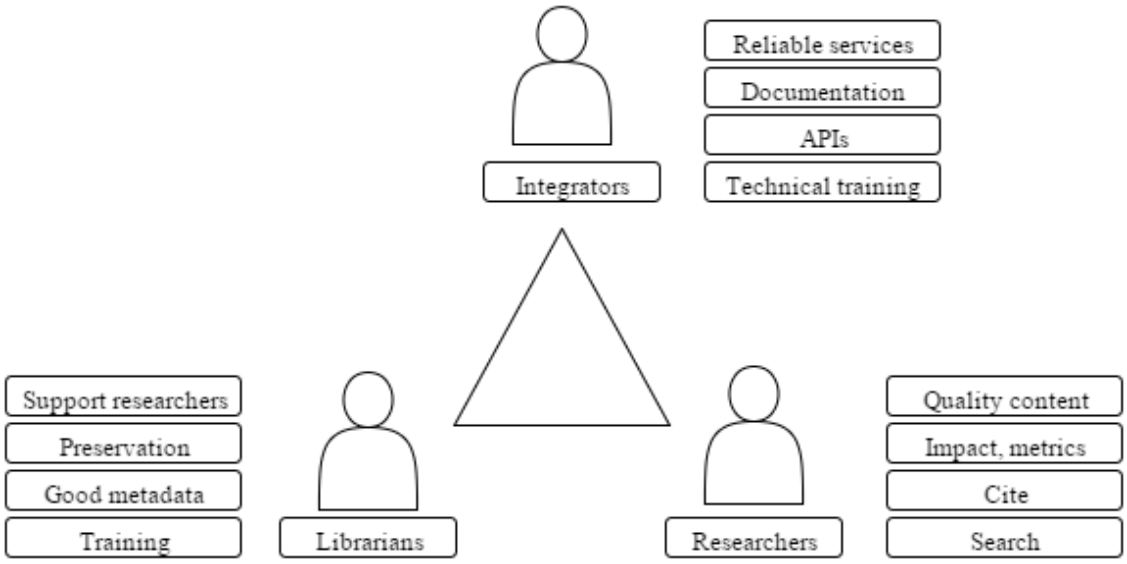

Figure 4. Communication audiences and their interests.

Integrators and developers are primarily interested in technical training. They want to learn about the different APIs, have a well documented and reliable infrastructure to build their services. For this group of stakeholders, we have improved our backend and developed tools to automate functionality. DataCite periodically hosts technical webinars and, through the THOR project, we are planning a series of face to face workshops, where we will be able to work together (Brown and Demeranville, 2015).

To reach the librarian stakeholder community, DataCite participates in multiple conferences geared for this community. We also collaborate in the development of standards and stay engaged in the community conversations. Our members are a key part of the development of DataCite's Metadata Standard and promote best practices. Social media provides an informal channel to interact with librarians, where we can discuss news and opportunities.

Engaging with the research community is much more challenging. It requires a scalable and diverse approach. DataCite's goal is to provide services that are transparent and easy for researchers to engage with, which includes a seamless infrastructure that does not require additional expertise details. To achieve this goal, we must work in close collaboration with DataCite's members, individual data centres, and libraries to simplify workflows, provide quality content, ensure data citation and impact tracking, and provide appropriate communication and outreach materials.

Conclusions

An interoperable e-infrastructure is an essential stepping stone for the establishment of data publication and reuse in the scholarly communication environment. The pivotal role of persistent identifiers eases the path, but still requires careful development, harmonised integrations and services, and a steady effort in communicating advantages and best practices.

DataCite has spent the last six years working towards a network of committed organisations and a reliable and transparent infrastructure. Through this process, we have understood the importance of consistent workflows in both the design and

development of integrations. Biased approaches – pushing for a model, neglecting the needs of the community, not interoperable or focusing on only on the technical level – have proven counterproductive. The community will only profit of the advantages of data publication if the services provided are solid, transparent for the final user and bound to standards.

Metadata quality is a critical issue. Although we have gotten far improving mismatches between different formats, it is still necessary to encourage and support data centres to provide comprehensive metadata. The most basic services, such as search or citation tracking, rely on the information provided by the data centre. We found diverse cases, but most of them can be improved with a better engagement of the researchers, providing them incentives and friendly submission processes, and stronger effort in metadata curation.

With the current level of integration, multiple initiatives are already working to develop training and outreach actions. DataCite is working with the THOR Project, tailoring actions for different stakeholders and their needs. This is one of the last steps to finalise the foundations of data publication. The adoption is still in its early stages and we have to keep on fostering it to make research data as essential as traditional publications in scholarly communication.

References

Brown, J., & Demeranville, T. (2015). D4.1 communications plan. THOR Project Deliverable.

Fenner, M., Demeranville, T., Kotarski, R., Vision, T., Rueda, L., Dasler, R., Haak, L., & Cruse, P. (2015). D.2.1: Contributor and organisation relationship data schema. THOR Project Deliverable. doi:10.5281/zenodo.30799

Kratz, J.E., & Strasser, C. (2015). Making data count. *Scientific Data*, *2*, 150039. doi:10.1038/sdata.2015.39

Rueda, L., & Dasler, R. (2015). D2.1: Artefact, contributor, and organisation relationship data schema: Appendix A. THOR Project Deliverable. doi:10.5281/zenodo.30800

Vision T., Kaye, J., & Aryani, A. (2014). Towards identifier-aware cyberinfrastructure for data and researchers. Paper presented at the 9th International Digital Curation Conference, San Francisco, USA. doi:10.6084/m9.figshare.907482