

Permanence of the Scholarly Record: Persistent Identification and Digital Preservation – A Roadmap

A. Dappert
The British Library
96 Euston Road, London, NW1 2DB
UK
angela.dappert@bl.uk
orcid.org/0000-0003-2614-6676

A. Farquhar
The British Library
96 Euston Road, London, NW1 2DB
UK
adam.farquhar@bl.uk
orcid.org/0000-0001-5331-6592

ABSTRACT

This paper proposes steps towards a roadmap for improving the integration of two communities that deal with persistence and long-term stewardship of digital content. They are Persistent Identifiers (PIDs) and Digital Preservation. Both disciplines have made significant progress and practical contributions. Yet their approaches are not fully linked and there is considerable potential to integrate their solution space and to improve either of them by learning from the other. It addresses three core issues:

1. How does the long-term digital object life-cycle affect PIDs, the entities they identify, and the metadata that describes them?
2. How can PIDs help long-term preservation?
3. How can long-term preservation help to shape PID best practice and ensure long-term access to the scholarly record?

We also sketch out initial results of our ongoing work along this roadmap.

KEYWORDS

Persistent identifiers, digital preservation, scholarly record, roadmap, distributed collaboration

1 INTRODUCTION

The Persistent Identifier and Digital Preservation communities both address core issues related to persistence and long-term stewardship of digital content. They have made significant progress and practical contributions over the past two decades. In spite of addressing shared challenges, the opportunities for each to benefit from the other's progress remain largely unrealized. This paper addresses three core questions to provide a roadmap to improve sharing of results:

1. How does the long-term digital object life-cycle affect PIDs, the entities they identify, and the metadata that describes them?
2. How can PIDs help long-term preservation?
3. How can long-term preservation help to shape PID best practice and ensure long-term access to the scholarly record?

While the work described can be applied to any digital material that is worth preserving, we emphasize the scholarly record. It is characterized by complex sets of contributors and long chains of information creation and exchange. These necessitate globally unique persistent identifiers more than many other digital materials.

1.1 Space, Time and Intent

Persistent identifiers (PIDs) play an important role in the scholarly infrastructure. They enable both people and computational agents to reliably identify and link to entities such as articles [e.g. provided by Crossref¹] or data [e.g. provided by DataCite²]. Furthermore, they can be used to identify researchers [e.g. provided by ORCID³] or rights-holders living or dead [e.g. provided by ISNI⁴]. PIDs are becoming an essential component in the workflows of funders, researchers, research organizations, data centers, publishers, libraries, and others. As of 2017, tens of millions of PIDs have been assigned to these core scholarly entities through global PID service providers and their partners.

Trusted and reliable identifiers associate a resource with a character string. The following *PID criteria* hold (inspired by [1]).

- A PID is a name, rather than an address.
- PIDs are globally unique.
- PIDs are persistent.⁵
- PIDs are selective at the right level of granularity.
- PIDs are interlinkable.
- PIDs are interoperable with other identifiers.
- PIDs are designed to last beyond the lifetime of any system or (most) organizations.
- PIDs are globally resolvable as a URI with support for the full range of HTTP including content negotiation.
- PIDs are managed through a sustained committed organization and governance process.

¹ <https://www.crossref.org/>

² <https://www.datacite.org/>

³ <https://orcid.org/>

⁴ <http://www.isni.org/>

⁵ This refers to the PID itself. It does not imply that the content must be persistent at all times. For example, the content may be streamed or be versioned.

- PIDs come with metadata that describes the resource's most relevant properties.

PID services mint PIDs on request and provide services such as registration, metadata management, fragment identification, content negotiation, search and discovery, and governance. Content owners manage the content and ensure that content location information is kept up-to-date with the PID service. Because both the PID service and the content owners hold metadata that describes the identified resources, PIDs can be indexed and searched. PID service providers, such as DataCite, establish a contractual governance commitment with content owners to ensure long-term stewardship and accessibility of the identified resource.

Leading motivations for PID use lie in the stewardship of the scholarly record. PIDs “improve the ease of locating resources; are actionable on the Web; enable metadata update and corrections without losing the resource's identity; can integrate legacy naming systems; promote linking and interoperability between services; and reduce confusion among versions of a resource. Widespread uptake of PID e-Infrastructures can accelerate the adoption of Open Science by building trust through seamless discovery of scientific artefacts; clear attribution to contributors; traceable provenance; unambiguous citation in scholarly discourse; supporting reproducibility; and enabling improved metadata quality through linking connected metadata sources.” [2].

PIDs can be applied to publications, data, other research outputs, researchers or other personas, organizations, legal entities, funding instruments, projects or patents, and more. They can also be used to distinguish aspects of an entity such as separate versions, multiple formats, levels of granularity; or of an object, such as its intellectual definition (e.g. a FRBR work or expression), or its rendition consisting of a bitstream, a single file, or a composite set of files.

PIDs and the relationships between them create a connected network of information about the global scholarly record. This is a graph in which the metadata of one PID relates to that of another

PID. PIDs can and are being used in the workflows of funders, researchers, research organizations, data centers, publishers, libraries, and others. Most of the discussion has been around their “contemporary” functions in the processing, use and reuse of resources, such as in data center or publisher workflows.

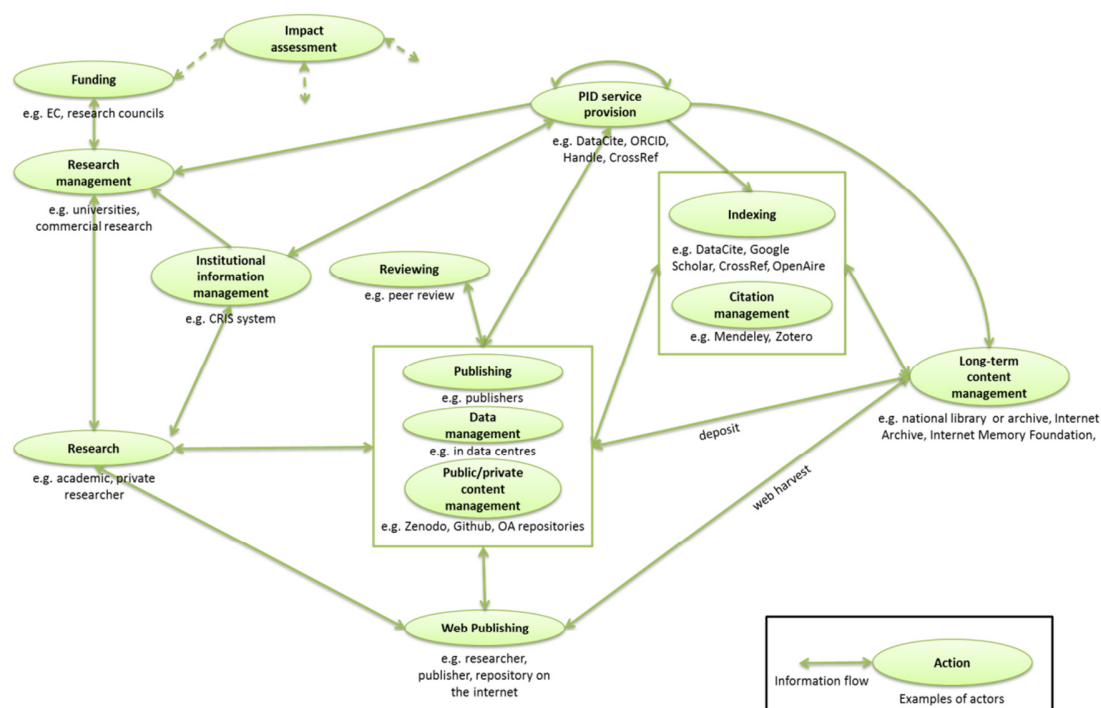
It is important to note that the scholarly record and this graph and the resources within it are used across three dimensions:

- Space: semantic linking of PID-identified resources creates the open eScience landscape in which we can globally connect and analyze data and associated metadata. This is emphasized in linked open data environments.
- Intent: Resources can be reused and re-purposed in ways that were unanticipated during their creation.
- Time: The life-cycle of research and the scholarly record spans centuries. It reaches from the conception of research ideas to reuse of results decades or centuries after their creation.

There has been much work towards creating the connected scholarly record and enabling validation and reuse of research outputs. In contrast, relatively little effort has gone towards ensuring long-term access to the scholarly record. It is this last dimension, “time”, for which this paper outlines a roadmap for future work.

Use cases across time include reserved PIDs for preliminary research outputs; transfer of responsibility for an entity from a creator or publisher to a memory institution; use of PIDs and content resolution after format migrations that are necessitated through obsolescence; handling deleted or lost data; PID creation for data that are created within memory institutions, when large data collections are mined resulting in derivative data sets; and provision of long-term stewardship for identified content.

Figure 1: PID-related information flow among stakeholders



We investigate the role that memory institutions and digital preservation practitioners play in the use and preservation of PIDs, their accompanying metadata, and the content they describe. Much of this is done by raising research questions whose answers will help improve both PID use and digital preservation practice.

2 THE SCHOLARLY RECORD LIFE-CYCLE

How does the digital object life-cycle affect PIDs, the entities they identify, and the metadata that describes them? To investigate this we need to look at the following issues:

1. What role should PID stakeholders play in order to ensure long-term preservation?
2. How can one manage the distributed long-term responsibility?
3. How can PIDs support entities that evolve over time?
4. How can we preserve the PID graph as it grows over time as more links are established through incremental improvements, use and reuse?

2.1 Role of stakeholders in ensuring long-term preservation of PID-related information

While PIDs are sometimes created for local use only, they are inherently about global reuse and establishing the connected scholarly record. Therefore, PID-related information is passed along from one organization to another across the life-cycle. They include:

PID services ⁶	Stakeholders
Funders	
Research organizations	
Researchers (as producers, consumers and reviewers)	
Publishers	
Data centers	
Institutional information managers	
Citation managers ⁷	
Indexing services ⁸	
Public content repositories ⁹	
Private content repositories ¹⁰	
Libraries and archives	

The Information Objects that are exchanged between these stakeholders are of three different types, with differing relations to the longevity of the scholarly record.

PIDs	Information Objects
Metadata	
Content	

⁶ ... that mint and manage PIDs and their associated metadata, such as DataCite, Handle, CrossRef

⁷ Such as Mendeley, Zotero

⁸ Such as Google Scholar, CrossRef, JISC KB+, SHERPA RoMEO, Directory of Open Access Journals (DOAJ), EBSCOHost.

⁹ Such as Zenodo, Github

¹⁰ Such as the Internet Archive

The stakeholders perform different Business Actions on the Information Objects over the life-cycle.

Funding	Business Actions
Research management	
Institutional information management	
Research	
Impact assessment	
Reviewing	
Publishing	
etc.	

Each business action consist of a sequence of basic actions that affect the Information Objects in the following ways.

Create	Basic Actions
Update	
Enrich	
Replace	
Delete	
Transform formats	
Disambiguate	
etc.	

Figure 1 shows an information network graph for PID-related information flow among the various business actions. Each node represents a class of business actions (e.g. PID service provision or provision of institutional information repository services); a link represents a flow of information objects (PIDs, metadata, or content). The nodes shown correspond to currently active actions; the links correspond to flows that are fairly well established as of 2017. Several actions may be performed by the same actor; e.g. a national library may provide PID services, run a web archive and perform long-term content management. A link implies that there is an information flow between some of the relevant actions covering some of the information objects. Reviewing this representation highlights links that are desirable for long-term preservation of the scholarly record, but may not be sufficient, or even exist, at the moment.

Where links are not well established, it is not possible for an actor to obtain information reliably or without substantial effort. Often links exist, but they do not pass along all of the valuable metadata that is associated with a PID. For example, in order for a library to acquire the information it needs about a dataset, it may be necessary to retrieve one subset from a data center, and another from a PID service provider.

A question is whether the information network has the links that are desirable in practice. For example, currently, there are no well-established links that inform Libraries and Archives about metadata updates in external PID Services, or that enables researchers to export PID-related metadata from Libraries and Archives into Citation Managers, or that pass rights metadata from PID Services to Researchers or from Data Centers to Libraries and Archives.

Even in the case that there are good links in the information network, it is not always clear who is responsible for assuring the long-term usability of an Information Object. To answer this, we

need to know how each organization type can best contribute to the assurance for long-term maintenance of the scholarly record; determine where in the life cycle it is easiest or most effective to create PIDs; determine where PIDs should be enriched with metadata; and when to perform other essential Actions.

2.1.1 Rules. As with all information creation, some rules-of-thumb apply. The closer information is collected to the source of creation, the easier it is to obtain and the more authority it holds. The longer-lived the custodial organization is, the more trust exists in its continued ability to support the information. The sooner in the life-cycle good housekeeping applies (such as assigning PIDs), the easier it is to avoid violations of the information on the way. The later in the life-cycle a stakeholder is positioned, the better is their ability to provide guidance on uniform formats and vocabularies and to grant comprehensive access. The later in the life-cycle a stakeholder is positioned, the more contradictory assertions may have accumulated about an identified entity and the more doubt there is on the provenance of metadata.

One step in the roadmap is to investigate these rules and which of the stakeholders should best be responsible for each Action / Information Object pair. For example, what metadata should be created by PID service providers to support long-term preservation? What content or metadata format transformations should data centers perform to support long-term preservation?

2.1.2 Longevity of organizations. Different stakeholders have different ideas of what long-term means. For example, time-limited organizations such as projects may produce web content with a life-span of months. Data centers have a considerably longer outlook, perhaps a decade or two. Institutional repositories are often private in nature and have no inherent incentive to provide links for capturing the overall scholarly record. While they may be long-term in nature, they do not have a long-term mandate, such as national cultural heritage institutions. Open Access repositories, such as Zenodo, assign PIDs and can function as an intermediary for content that has no other organizational PID support, but may not have the long-term mandate of a national cultural institution. Similarly, privately held organizations dedicated to long-term access, such as the Internet Archive, may be more vulnerable than public institutions such as national libraries and archives that are backed by a national commitment.

National memory institutions have a fundamental long-term mandate and can offer a safety net for valuable cultural and scientific assets. This applies to the metadata and content associated with PIDs; it also applies to content that is not otherwise eligible for PIDs because it does not have a long-term home. What new roles should these organizations play in the long-term preservation of the scholarly record? Who takes responsibility for creating PIDs and for creating metadata that is associated with the PID?

Considering the information network and the varying long-term commitment of the stakeholders in the scholarly record, there are two crucial questions to address:

- How should the handover of Information Objects between stakeholders be governed and managed?
- Who takes responsibility for minting PIDs?

2.1.3 Handover. There are many pragmatic questions that must be addressed to ensure that the handover of *Information Objects* between stakeholders is effective. These include:

- Where should business-as-usual handover of any information related to the scholarly record be initiated? Should handovers be defined and governed in a systematic manner? How can business models of different organization types define a hand-off best practice, as we are already familiar with from, for example, the hand-over between records management systems to archives? How is responsibility to be transferred technically?
- What should happen when preservation trigger events occur? As a proactive example, online digital research repository Figshare¹¹ has joined the Digital Preservation Network (DPN)¹². Their announcement states that “Research data made public on Figshare will be deposited into DPN, a dark archive that preserves scholarship for future generations. Figshare users can guarantee that long-term access to their scholarly resources will be protected in the event of any type of change in administrative or physical institutional environments.” Similarly, some journal publishers subscribe to the CLOCKSS dark archiving system to protect against the case that they may no longer be able to make their outputs available. A more reactive example of a sustainable governance migration is the handover of the PURL¹³ PID service management from OCLC¹⁴ to the Internet Archive¹⁵, which has a declared long-term business model¹⁶. In the event of organizational failure, there should be mechanisms in place for both the identifiers and the identified entities with a governance structure and a method to offer resolution and access services.
- Who should be responsible for aggregating the distributed scholarly record? PID services collect metadata on identified and related entities. This may give the impression that they guarantee the long-term availability of the scholarly record. But PID services have limited scope; their mandate only extends to the persistence of the identifiers, discovery metadata, and resolution services. For example ORCID’s¹⁷ main goal is to provide PIDs for researchers. ORCID also collects information about these researchers’ scholarly output such as alternative person identifiers, histories, funding, patents, and associated works. But it is not clear that it is ORCID’s responsibility to guarantee this metadata for the

¹¹ <https://figshare.com/>

¹² <http://duraspace.org/articles/2769>

¹³ <http://www.archive.org/services/purl>

¹⁴ <http://www.oclc.org/>

¹⁵ <https://archive.org/>

¹⁶ <https://www.oclc.org/en-UK/news/releases/2016/201623dublin.html>

¹⁷ <https://orcid.org/>

long-term. Where is the right scope for each of these PID services?

- Conversely, how do we ensure that the PID graph is fully connected where possible so that there are no unintended islands of information?
- What is the role and responsibility of a national library in preservation of globally distributed metadata that is associated with various PIDs across multiple independent providers?
- Metadata associated with content is often deleted upon handover between stakeholders. This may be for very good reasons. For example, when an image is shared on the web, one may wish to remove identifying information to comply with data protection regulations. If there are multiple copies of some content that is identified by a PID, there may be different metadata associated with each copy; how can a digital object consumer identify which copy holds the metadata they need or are entitled to?

2.1.4 PID minting responsibility. Organizations can only mint PIDs if they make a long-term commitment to enable access to the identified content. But there is valuable scholarly content, for example in the form of blogs, that has no dedicated long-term champion.

Webarchives in public and private content repositories and in national libraries and archives can provide persistence for some of these digital assets that do not have owners who can commit to their long-term accessibility. An interesting proposal by Zierau et al [3, 4] bases these assets' persistent identification on the persistent identification of the web archive itself. A PWID consisting of an identifier for the web archive, the harvest date-time, the harvested URL, and the context specification permits persistent global identification of any harvested web content with the guarantee of permanence provided by the webarchive, rather than the content originator.

2.2 Managing distributed long-term responsibility

No single organization today holds even a copy of the full scholarly record including PIDs, associated metadata, and the identified content, much less holds an authoritative copy. While concentration of information can simplify large-scale use, it may increase the risk of large-scale loss. Distribution and redundancy offer a form of resilience and improved availability. Given this decentralization of stewardship, how can one manage the distributed long-term responsibility? How can one avoid discontinuities in modeling and interfaces, to ensure interoperability at the edges of organizations' scope?

The Scholix¹⁸ framework offers a conceptual model, an information model, information standards and encoding guidelines, and options for exchange protocols toward solving interoperability issues. It is "a high level interoperability

framework for exchanging information about the links between scholarly literature and data". But there is also a need for technical, governance, and coordination solutions, in particular for providing long-term availability of the scientific record without requiring a central uniform repository.

If this information is distributed over the global web and exposed through shared protocols, it is available for dynamic harvesting for as long as the information is available on the web. But we know that web sources appear and disappear at alarming rates. There is a natural role for long-term stewards, such as national memory institutions, to harvest, preserve, and potentially provide access to the scholarly record.

For long-term stewards of information, questions arise as to the best models for managing PIDs that have been minted by multiple sources that possibly duplicate or overlap; or for content that has been combined from multiple sources.

As research advances, the scholarly record keeps growing. Over time, more links are established through incremental improvements, use and reuse. How can comprehensive harvesting of the scholarly record be assured for this? How can completeness be assured? How can one deal with contradictory sources of evidence?

2.3 How can PIDs best be used to support entities that evolve over time?

The scholarly life-cycle involves entities that are evolving over time. There may be changes to metadata or content. When, for example, datasets and software identified by the PID service change or related versions are created, one must track how each version relates to earlier ones. Each version may be identified by a new PID and linked through meaningful relationship types.

Whether or not changes establish a substantially different object that deserves the assignment of a new PID depends on the use case that is supported by it, and on the policies that underline the use case. For example, changing the spelling on an author's name may be considered a minor correction that does not necessitate the assignment of a new PID to a book. There is no use case that would handle the corrected object differently. Even so, one can keep a cumulative trace of any corrections. Adding an author to a book may necessitate the creation of a newly identified object, because it may be necessary to reflect the fact that different copyright assumptions were made before and after the correction. That is to say, there is a use case that results in different actions on the two identified objects. Therefore, one should distinguish the corrected object from the earlier one through a new PID and one should record the relationship between them, as well as the event and the policy that necessitated the creation of the new PID.

As a consequence, it is not the PID service that determines at what level of granularity PIDs should be assigned. PIDs support the clients' use cases. The PID services have to flexibly accommodate different client policies and use cases.

But this also implies that PID service providers have little control over the resulting granularity of the research objects that are identified. It is then the PID minting clients that can negotiate

¹⁸ <http://www.scholix.org/>

among each other to establish guidelines on policies that both support the implemented use cases and support interoperability and information exchange between different institutions.

These dynamics don't only apply to defining what use cases trigger versioning, but also to controlled vocabularies or the granularity at which PIDs are assigned to digital objects. For example, different even types necessitate the creation of a new PID for a derivative dataset. To support long-term management of evolving research data one would want to record the dataset's provenance by recording the relevant events. These events types are typically defined by a controlled vocabulary, such as "software patch applied", "time-filter applied". The PID service can define suggested controlled vocabularies, but the client must be able to use their own personalized vocabulary to meet their individual use cases.

One of the key R&D questions is, therefore, what functionality needs to be provided by PID services to enable their clients to capture the necessary versioning information about evolving entities. Initial discussions can be found in [5].

Memory institutions are practiced in dealing with these questions in the context of their digital repositories. Digital preservation metadata work, as discussed in Section 4 has provided recommendations for how to handle these situations that now can be applied to new contexts, such as PID services.

3 MEMORY INSTITUTIONS - PIDs HELP LONG-TERM PRESERVATION

Using PIDs can improve processes for institutions that need to satisfy a long-term mandate.

3.1 Authority control

Wikipedia [6] states that in "library science, authority control is a process that organizes bibliographic information, for example in library catalogs by using a single, distinct spelling of a name (heading) or a numeric identifier for each topic. ... These one-of-a-kind headings or identifiers are applied consistently throughout catalogs which make use of the respective authority file, and are applied for other methods of organizing data such as linkages and cross references." Authority control supports information management and is shifting toward PID-based rather than string-based solutions. This is a significant change in the information management practices in memory institutions that helps to avoid shortfalls of string-based authority practices, such as spelling errors. For example, using an ISNI for current or historic rights holders or an ORCID for self-registered researchers supports precise unambiguous identification in citation, information linkage, or even authentication.

The roadmap needs to identify:

- What PID functionality is needed to transition from string-based authority or local identifiers to globally unique PIDs?
- What adjustments are needed in memory institutions' workflows to transition to PID-based authority control?

- What adjustments are needed in PID services and governance to meet the quality, scalability, affordability, and other requirements of memory institutions?

3.2 Infrastructure for preservation of the digital scholarly record

Digital preservation is a form of long-term information management. Institutional repositories can provide local storage and archiving for scholarly outputs and they are equipped to manage metadata and some content. Realistically, the ability to preserve the wide array of research object types, such as non-SQL databases, is limited and needs further work. Furthermore, repositories are not equipped to provide content resolution services to replace failed PID services. The practice in PID services can inform improvements in digital preservation services in this regard. And, as mentioned earlier, repositories have limitations in managing PID related metadata when they are not the organization on record and known to the service that minted the PID.

If memory institutions harvest content from the web, PIDs can be very helpful, but still have implementation inconsistencies that prevent effective automated harvesting. We need to improve managing long-term data and artifacts that include PIDs so that they can be used to better streamline digital preservation efforts. For an example, see Van de Sompel, Rosenthal, Nelson's [7] discussion on eJournal preservation.

3.3 PID use in digital preservation repositories

In digital preservation, identification of digital content is essential. In the contemporary scholarly process, PIDs are used for validation and reuse of research results. Long-term reuse of material that is held by memory institutions is even more challenging, since the material is created and consumed by third parties. Reuse happens over much longer periods of time. As a result assumptions about the environment and context required are less likely to hold. This makes PID use even more important. But by far not all content in digital long-term repositories is identified through PIDs. Most digital repositories deal with a variety of identifiers, most of which lack one or more of the *PID criteria* outlined above. It is essential to understand how long-term repositories can enhance existing local or transient identifiers within their scope to support the PID criteria.

PREMIS states [8] that "for a given identifier to be usable, it is necessary to know the identifier scheme and the namespace in which it is unique. If a particular repository uses only one type of identifier, the repository would not need to record the scheme in association with each object. The repository would, however, need to know this information and to be able to supply it when exchanging metadata with other repositories." This requirement only ensures that an entity is identifiable within one repository. To support a global information network, it would require very precise knowledge about when a scheme was applied, how the scheme changed over time, how the versions relate, and so on. As Information Objects flow through the information network, chances are that a long-term repository would not be able to

collect the essential information in external schemes to establish identifiability. The question to address is how PID use can support this information flow.

Another aspect of long-term information management is a need for heightened resilience. The longer-lived content is, the more likely it is that parts of the information object may inadvertently be corrupted. For example, we have anecdotally witnessed that the links between PIDs and their associated metadata have been broken through software programming errors so that PIDs could no longer be linked to the content they identified. In order to mitigate this sort of risk it is advisable to not solely rely on PIDs, but to judgmentally enrich them with redundant metadata that would permit a semantic match of entities between distributed long-term systems if the PIDs themselves get corrupted. The questions that arise are, what metadata is best suited to serve this purpose, and how best to ensure synchronization of redundant metadata in multiple places.

3.4 Other archival tasks that can benefit from PID use

Many more entities are managed in memory institutions' information governance. Ideally there would not only be PID services for research outputs and agents (persons and organizations), but also for funders, grants, laws, patents, software packages, events, etc. Every PID service requires a governance structure, a metadata scheme, and support for information creation and exchange. These are currently missing for many entity types. Memory institutions are now also creating new and derived data sets related to their digital collections through text and data mining, analysis, crowd sourcing, and citizen science. Stable ways of identifying these data sets, their provenance and their contributors need to be implemented.

4 PID SERVICES - LONG-TERM PRESERVATION HELPS TO SHAPE PID PRACTICE

The digital preservation community has developed an array of practices and techniques for long-term information management. How can this help to shape PID best practice and ensure long-term access to the scholarly record? Where are gaps in PID services' current practice that can be informed by memory institutions' practice?

A key answer is the metadata that is associated with PIDs and the entities that are identified by them. PREMIS [8] is the de facto metadata standard for long-term access to digital content. It recommends the information about digital content that is very likely needed for long-term use and preservation. Its goal is to ensure the availability, identifiability, integrity, viability, renderability, understandability and authenticity of digital content. PREMIS articulates data modeling and metadata principles that should be adopted early to remove vulnerabilities for digital content and to ensure its long-term usability. But PREMIS does not just address digital preservation. A file can become unreadable

because its format has become obsolete over years or because there was a power failure during file transfer today – both situations require similar consideration. Therefore, the approach taken in PREMIS can be helpful for near-term management of digital content.

Figure 1 illustrated how metadata that is associated with the scholarly record is passed through a network of stakeholders' actions until it ends up in memory institutions for long-term content management. The stakeholders involved in the scholarly information network can benefit from those PREMIS principles. For example, PID services, such as CrossRef or DataCite, are justified in specializing on content types, such as scientific articles, monographs, or data sets. As their portfolios grow to accommodate newly supported types, their data models need to be extended. The first-principles approach taken in PREMIS helps to ensure that data models are both extensible and interoperable. Since PREMIS is expressed as framework and principles it can be implemented in any implementation environment.

Future work should articulate those recommended PREMIS features, perform a gap analysis to understand to what degree they are currently not supported with scholarly stakeholders, what implications this has for the long-term access to the scholarly record, and how these short-comings could be overcome.

An initial set of example recommendations of how stakeholders in the scholarly information network can adopt good practice from PREMIS is discussed in the following.

PREMIS distinguishes Objects, Actors (including, people, organizations and computing components), Rights statements and Events, which are sufficient to support modeling most semantic relationships required. Objects can be described on four separate, declared levels:

- the intellectual description that helps search and find the object;
- representations, which are sets of files that together create one rendition or execution of a digital object;
- component files of each representation;
- and individual bitstreams.

Efforts that model digital objects without distinguishing these levels often end up confusing issues that should be separated. For example, some PID solutions conflate resolution to a landing page that holds descriptive information (similar to the intellectual entity) with resolution to the content itself (which could be a representation, file, or bitstream). This lack of clarity hinders automatic crawls, machine-harvesting and machine-interpretation of parts of the scholarly record. Van de Sompe et al. [9, 10] analyze this situation for web use of PIDs. Adopting the PREMIS model would resolve this challenge. Different users may need to resolve to different levels. For example, a crawl bot may just want to find representations and files; a search indexer may just be interested in intellectual entities and descriptive metadata; someone running an impact analysis may just want to identify the researchers and, from there, link to their research outputs. Because of these varying use cases and goals the underlying conceptualizations need to clearly identify the types of entities of

interest. This results in access mechanisms that support these use cases flexibly. PREMIS entities are sufficient to support modeling most semantic relationships required to capture these distinctions. Technical metadata, as specified for PREMIS objects, should be created as early as possible in the information network (e.g., including information about file types, checksums, creating software, or computing platform requirements).

The PREMIS model covers derived, dependent or structurally related objects and provides a transparent way of relating them with each other. Each relationship can document the nature of the relationship, and the events and agents that were involved in creating it. As we have seen in Section 2.3, this ability is important to support use cases for many of the actors that rely on PID services.

PREMIS provides two powerful tools for identifying and describing partial and dynamic datasets. The first one is the ability to describe and identify bitstreams (mentioned above). If data fragments can be described as sets of bitstreams, they can be directly identified and described.

Sometimes they are, however, better described by the event that created a dynamically derived subset (e.g., for data sets that are derived from a base data set by applying a filter). This can be addressed by events associated with relationships. The derived dataset can be identified by annotating its relationship to the original data set with an event that captures the selection criteria or selection algorithm. The derived data set is identified, but does not actually need to be instantiated; it can be computed on demand from the original data set. In this way PREMIS modeling can be used to support the implementation of the RDA recommendations for Data Citation of Evolving Data [11] within its basic inherent data model and without the need for creating any extended functionality for special cases.

Relationships also let you relate data sets that are related through a structural hierarchical inclusion relationship on various levels of granularity. PREMIS objects can be used to describe software and hardware or other parts of the computation stack, which are essential components in the scholarly record that are currently not adequately covered by PID services. Adopting PREMIS compatible conceptualizations would make it easier to provide PID support to these object types and to support interoperability among scholarly stakeholders.

Another example is the maintenance of provenance information in the scholarly record. Provenance may contain events from any point in the object life cycle. For example, they might record due diligence activities, or events that transform an object or its metadata. It is important to record these events to determine the degree of authenticity of the object over time.

PREMIS events also encourage linking to event-related agents, ranging from data creators to researchers, curators, to publishers, but also organizations or software agents. This form of modeling does not simply identify the role of an agent with respect to an object, as is currently frequently done, but it specifies in which event this role was taken. This allows for a much more precise and time-linked recording of the agents' roles.

CrossRef and DataCite are starting to collect and distribute information about events related to PIDs. For example, when an article cites a dataset or when a new version of a dataset is released. Currently, this is not based on a generalizable event model, and it does not extend to provenance events. The responsibility for this sort of information may mainly lie with their clients, but PID service providers should consider collecting it as additional assurance. Again, adopting the PREMIS modeling style for events (or agents or rights) can improve interoperability across the information network and simplify information exchange.

These examples illustrate some of the PREMIS features whose adoption could improve the information exchange in the scholarly information network. Future work may investigate how these principles can best be translated to individual implementation environments.

5 CONCLUSIONS

We set out to investigate how Persistent Identifier services can be extended to long-term information management. When starting to address this, we realized that the number of unsolved issues was substantial and should not be addressed ad hoc. A more systematic analysis of the research and development space was required to combine lessons from the two domains. Consequently, this paper sets out key questions and challenges for a roadmap to improve the alignment between Persistent Identifier and Digital Preservation approaches. Due to the real advances that have been made in each community, this alignment may enable a more deliberate design rather than stepping through ad-hoc improvements. Enhanced aligned services involved in creating and maintaining the scholarly record would support a more complete information flow. The resulting data models would explicitly support long-term preservation of PID resolution services, metadata that captures information about the scholarly record, as well as access to the actual research objects.

ACKNOWLEDGEMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654039". We would like to thank our reviewers Herbert Van de Sompel, Ignasi Labastida, and Barbara Sierman for suggesting work in this area.

REFERENCES

- [1] Consortium, ODIN; Fenner, Martin; Thorisson, Gudmundur; Ruiz, Sergio; Brase, Jan (2013): D4.1 Conceptual model of interoperability. figshare. <https://doi.org/10.6084/m9.figshare.824314.v1>
- [2] A. Dappert, A. Farquhar, R. Kotarski, & K. Hewlett (2017). Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research. Data Science Journal, 16, 28. <http://doi.org/10.5334/dsj-2017-028>
- [3] E. Zierau, C Nyvang, T Hvid Kromann (2016) Persistent Web References – Best Practices and New Suggestions, iPRES 2016, 13th International Conference on Digital Preservation
- [4] C. Nyvang, T. Hvid Kromann, E. Zierau (2017) Capturing the web at large - A critique of current web citation practices. Conference Researchers, practitioners and their use of the archived web. 14 Jun 2017, https://archivedweb.blogs.sas.ac.uk/files/2017/06/RESAW2017-NyvangKromannZierau-Capturing_the_web_at_large.pdf. Accessed 17 July 2017
- [5] M. Fenner, T. Demeranville, R. Kotarski, R. Dasler, J. McEntyre, G. de Mello, T. Vision, A. Dappert, A. Farquhar. (2016). THOR: Conceptual Model of Persistent Identifier Linking. Zenodo. <https://doi.org/10.5281/zenodo.48705>
- [6] Wikipedia (2017). Accessed on 30 March 2017 https://en.wikipedia.org/w/index.php?title=Authority_control&oldid=764676731
- [7] Herbert Van de Sompel, David S. H. Rosenthal, Michael L. Nelson. Web Infrastructure to Support e-Journal Preservation (and More) (Submitted on 19 May 2016), [arXiv:1605.06154v1](https://arxiv.org/abs/1605.06154v1)
- [8] PREMIS Editorial Committee (2015). PREMIS Data Dictionary for Preservation Metadata, Version 3.0. <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf> . Accessed 17 July 2017
- [9] Herbert Van de Sompel, Martin Klein, Shawn M. Jones (2016) Persistent URIs Must Be Used To Be Persistent arXiv:1602.09102
- [10] Identifier. Signposting the Scholarly Web. <http://signposting.org/identifier> . Accessed 17 July 2017
- [11] Rauber, Andreas, Asmi, Ari, van Uytvanck, Dieter and Stefan Pröll. (2015). Data Citation of Evolving Data: Recommendations of the RDA Working Group on Data Citation (WGDC). Research Data Alliance. Retrieved from: https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf . Accessed 17 July 2017