Scholarly Resource Linking: Building out a "Relationship Life Cycle"

Matthew S. Mayernik
National Center for Atmospheric Research (NCAR),
University Corporation for Atmospheric Research (UCAR),
USA. mayernik@ucar.edu

ABSTRACT

Scholarly resources, including publications, software, data sets, and instruments, are created in an iterative and interrelated fashion. Managing the relationships that exist among and between such resources is a central requirement for information systems. Practically, however, many scholarly resources exist online as discrete entities, divorced from other resources to which they are intimately related. A robust system for linking scholarly resources in a broad and sustainable fashion will have to navigate a set of complex and interrelated requirements. This paper presents results and insights from three different projects that focused on supporting more robust linkages among scholarly resources. The discussion details key technical and institutional challenges looking forward and backward in time across what might be considered to be a "relationship life cycle": identifying, validating, characterizing, and preserving relationships. The goal of the paper is to help guide new research initiatives and operational services focused on integrating relationship information into the scholarly record.

KEYWORDS

Data citation, metadata, web linking, scholarly communication

INTRODUCTION

Most scholarly resources are now available online. Journal articles are published via online platforms, reports and other gray literature are available via institutional and general-purpose repositories, data sets are increasingly archived in web-accessible data repositories, and software packages are widely distributed through GitHub and other code sharing tools. Not all resources are available online, of course. Numerous well-known social, cultural, and technical factors impede distribution of research products, and scholars do in some cases ensure a competitive advantage by maximizing their unique access to novel data, tools, or knowledge (Mitroff, 1974; Fienberg, Martin, & Straf, 1985; Borgman, 2012). It is certainly also the case that making something available online does not automatically ensure its usefulness or understandability to a broad audience (Mayernik, 2017). Nonetheless, the trajectory is clear that scholarly resources already are, or will increasingly become, available online. Policy pressures, technical affordances, and social norms are all pushing in this direction (Willinsky, 2005; Woelfle, Olliaro, & Todd, 2011; Kriesberg, et al., 2017). In the context of scholarly work, having resources available online provides broad benefits. Scholarly communities benefit from increased access to large numbers of resources, and individual scholars benefit from the citation and readership advantages that accrue (Hitchcock, 2013; Piwowar & Vision, 2013).

Once resources are available online, a natural question arises: can we link resources together? Linking, after all, is what the internet is all about. Publications, software, data sets, and other scholarly resources are created in an iterative and interrelated fashion. The interconnections between scholarly resources can be characterized as forming a value chain in which relationships provide significant value for resource discovery, use, management, and preservation approaches (Van de Sompel, et al., 2004; Pepe, et al., 2010). Managing such relationships is, in principle, a central component of information systems (Kent, 1978). Indeed, many information and data systems do manage and leverage relationships of a variety of kinds, particularly relationships among vocabulary terms and content structures (Bean & Green, 2001). Practically, however, most scholarly resources exist online as discrete entities. A search of the Registry of Research Data Repositories (https://www.re3data.org, search done on March 28, 2018) found only 48 repositories out of 2040 total repositories that are identified as accepting "Source Code" and 26 repositories as accepting "Software applications," with some repositories accepting both. These numbers are likely not exact, but they illustrate how data and software repository services are largely disjoint. Similarly, scholarly papers are housed in systems and repositories that typically do not collect data or software assets.

Linking scholarly resources thus requires an approach that navigates multiple scholarly institutions and technical systems. Establishing and managing relationships between information resources has been a common theme in information science and technology research. This paper will not attempt to survey all of the relevant literature and initiatives in this space. Useful overviews can be found elsewhere (Borgman, 2007; Lagoze, 2010; Van de Sompel & Nelson, 2015; Mayernik, Phillips, & Nienhouse, 2016). Instead, this paper presents results and insights from three different projects focused on linking data and literature. It outlines key challenges in identifying, validating, characterizing, and preserving relationships among scholarly resources, and discusses how these challenges differ when looking forward or backward in time. The goal of the paper is to help guide new research and operational services focused on integrating relationship information more fully into the scholarly record.

INVESTIGATIONS INTO LINKING

This section presents three separate projects conducted in the past four years. Each project has had different specific goals, scopes, and partners, but the overarching theme among the three projects has been to investigate approaches to cross-linking related scholarly resources. As Van de Sompel and Nelson (2015) state in their own recent retrospective on a series of scholarly resource and infrastructure interoperability projects, "it is hard to know how to exactly start an effort to work towards increased interoperability," because of the diversity of challenges and stakeholders involved. The following three projects should be seen as an attempt to make multiple starts in parallel toward understanding and potentially addressing scholarly resource cross-linking challenges.

Below, each project is discussed individually. Each section describes the projects' respective goals, approaches, and relevant accomplishments. Each section also identifies some notable lessons learned via the work within each project. Following these project descriptions, I provide cross-cutting discussion that pulls together points from each project to develop broadly applicable insights for scholarly resource cross-linking efforts.

Repository-to-Repository Cross-Linking

The first project (active Jan. 2015 – Mar. 2016) consisted of a pilot effort to exchange link information between two repositories, one data repository and one literature repository. The goal was to enable the two repositories to interact directly to exchange link information for resources that had known relations, but were hosted and managed separately, such as, for example, if data sets hosted by repository A were used to produce publications that were hosted by repository B. The vision for the project was to develop an interchange process between the two repositories that allowed researchers who deposited resources in one system to initiate the deposit of related resources in the other. Ideally, links between related resources held in the two separate systems could be exchanged and made visible in the respective repositories. From a system perspective, the objectives were to identify, scope, test, and deploy repository features that allow links between related resources to be created, exchanged, and maintained over time with low technical hurdles and minimal repository curator effort. The full end-to-end implementation of this vision was not completed, as discussed below. But the process of working on the project enabled us to investigate key requirements for how repository cross-linking might be achieved. We also held a workshop at the end of the project in which broader discussion of the project topics took place (Mayernik, Phillips, & Nienhouse, 2016).

We approached the project by: 1) specifying use cases and stakeholders, 2) developing technical requirements for these use cases and stakeholders, and 3) developing system functionalities that could meet the use case and stakeholder requirements. Figure 1 shows four resource linking scenarios that relate to the use cases and stakeholders for connecting a document with its underlying data. In scenarios #1-3 in Figure 1, a content creator (e.g. publication and/or data author) might be interested in depositing two new resources or might be depositing a new resource that is related to a resource deposited previously. In scenario #4, repository curators might be interested in identifying links between resources that are already part of their existing collections.

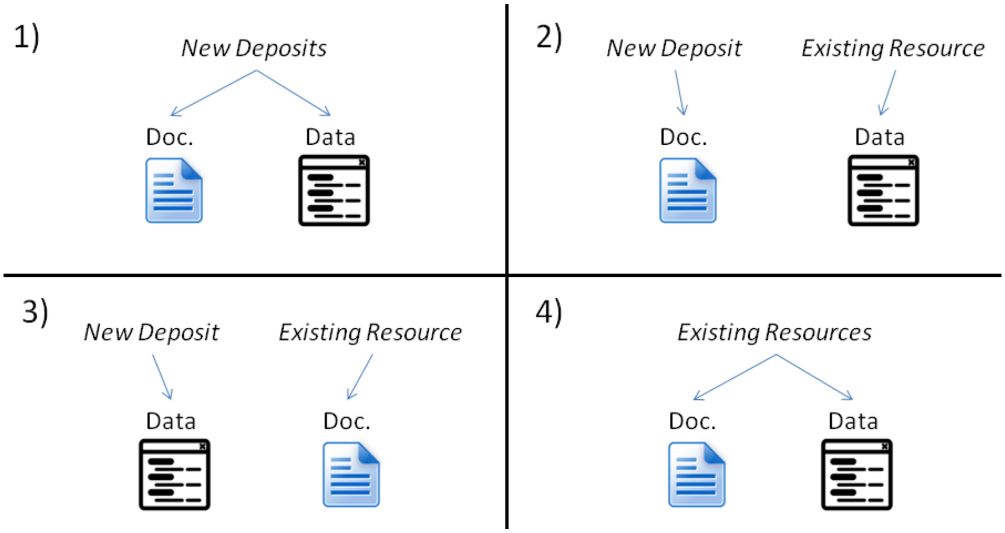

Figure 1. Repository linking workflow scenarios

These scenarios presented a number of tasks to support, including that: a) Data set providers can upload publications into one repository when depositing data sets in another (or vice versa), b) Repository content curators would be notified when new submissions are initiated from an external system, c) Publication and data authors would be able to share the deposited resources with colleagues and reference them appropriately, d) Research output consumers (either human or machine) could use the repositories to find resources and their relations. In breaking these tasks into sub-tasks and distinct workflows, the following technical requirements were identified being as key to establishing and maintaining robust connections between repositories of related resources:

- A notification mechanism to send information about related resources between repositories
- A data model for metadata exchange, to:
    - Specify metadata inheritance (e.g. authors and affiliated organizations that are in common from one resource to another)
    - Specify relationship types (e.g. that a data set underlies a publication)
    - Specify information necessary for maintaining relationships between resources over time as updates occur (e.g. a new version of a data set is produced, or a new publication is created based on an existing data set)
- Effective resource deposit interfaces that allow simple declaration of cross-repository relationships between data and publications
- Effective interface displays that make the cross-repository relationships transparent to the users

We investigated the SHARE Notify (https://share.osf.io/) service as a notification tool to enable repositories to interact and exchange relationship information. At the time of our investigation, SHARE Notify was being developed via the Center for Open Science as a third-party service for creating notifications of research publication “events.” Our interest in using SHARE Notify was to send notifications of relationship "events" between our repositories, with the SHARE Notify service sitting as a intermediary between the two repositories. The appeal of a third-party intermediary like SHARE Notify was that it eliminated the need for every repository to directly connect to every other repository. While our direct use case was to connect two repositories, we were interested in a means of distributing relationship information to numerous potential partners. Thus, we wanted an approach that could potentially scale to support interchanges of relationship information between many repositories. We were hoping to find a third-party that could provide central registry of research release “events” via a metadata feed. The SHARE Notify service had potential to support our desired features, though our use case was not its main focus.

We conducted pilot experiments to use the SHARE Notify service, but ultimately did not complete an end-to-end workflow during the timeline of our project, for a couple of reasons. First, the SHARE Notify service was itself in a phase of development and iteration. As such, it presented a moving target for our requirements. The SHARE Notify data model and APIs, for example, were not finalized during the period in which our project was active. This is a key lesson from the project. The stability of a third-party notification service is an external factor that can affect the reliability of local workflows.

Another lesson learned, not related to SHARE Notify, was that an effort such as ours has to be agnostic about the identification schemes used to identify resources. Some of the technical workflows we explored required assets to be identified via Digital Object Identifiers (DOIs), specifically to take advantage of the metadata stores hosted by DOI registration agencies (e.g. CrossRef and DataCite). Many resources in the two repositories of interest for this project, however, had not been (and were not expected to be) assigned DOIs.

Tracing Resource Linkages through Literature References

The second project featured here (active 2014-current) focused on developing and evaluating tools for automating the tracing of research infrastructures in the research literature via persistent citable identifiers. From a linking perspective, the interest was in tracing linkages between papers and associated resources, such as data sets, that were used to produce the papers. This project was motivated by the increase in interest in “data citation,” that is, the assignment of persistent identifiers like Digital Object Identifiers (DOIs) to data with the goal of enabling them to be cited and tracked like traditional research literature (Mayernik, et al, 2017; Silvello, 2017). Our interest, however, was in the question of the utility of persistent identification for research resources more broadly, including for scientific equipment (or sets of instruments), scientific software packages, computing systems, and communication networks. We use the overarching term “research infrastructures” to encompass this broad grouping. Since assigning and using persistent identifiers for scientific research infrastructures is a relatively new development, very few assessments have been conducted that systematically examine the effects of such identifiers assigned to these infrastructures. Our project goals were thus 1) to develop an understanding of how to methodically and consistently analyze the scientific impact of research infrastructures, and 2) to develop automated techniques that enable the tracing of research infrastructures to work more effectively. Specific research questions focused, for example, on how references to scientific research infrastructures have changed over time in relation to the assignment of persistent identifiers to data, software, and other components of research infrastructures.

The first effort within the project was to conduct a case study-based assessment to evaluate whether research infrastructures are being increasingly identified and referenced in the research literature via persistent citable identifiers. In this study (Mayernik & Maull, 2017), citations and references for four resources - two data sets, a software package, and a supercomputing facility - were collected manually using Google Scholar searches and analyzed to assess the ways that the resources had been referenced by researchers who used them, by characterizing how often the resources were referenced in papers through in-text descriptions, mentioned via acknowledgements, or explicitly cited in reference lists. Findings from this study showed how persistent identifiers assigned to the four examined resources were indeed being used in references in published papers on an increasing basis. But there was not a consistent pattern across the four case studies of how these increases were manifesting. Likewise, this analysis found that referencing practices were changing over time, but the extent of the changes varies considerably from resource to resource. A key takeaway from these results is that changing established practices for referencing and acknowledging data sets will potentially be more difficult than creating new practices for referencing other kinds of products, like software or computing facilities.

The above study was done as a largely manual process. The second effort within this project was to develop computational algorithms/methods to make this kind of assessment project easier. The development of tracing algorithms and methods followed a typical machine learning approach, centered on three axes: (1) collection of candidate publications for classification, (2) development of an experimental methodology for classification, and (3) an automation framework for executing document classification and analysis. This approach is showing promise, with an initial test of document classifiers being able to correctly determine whether a document from the test set did in fact use a computation facility of interest, based on characteristics of the documents' metadata and full-text. The limitation of this study thus far is that it is based on a relatively small corpus of documents. The training set (those that have been manually examined and labeled) includes about 300 documents, and the test set (those which had been labeled but had not yet been seen or trained on) includes about 120 documents. These numbers are far from ideal for machine learning studies, which often use thousands or even millions of documents.

A key lesson of this project is that human expertise is hard to scale but is critically necessary to identify and validate relationships in cases where computation-based approaches are not possible. Machine automation is readily scalable but ensuring and measuring accuracy of automated relationship gathering tools has been very difficult.

Both of the studies undertaken within this project have faced a similar challenge, namely that gathering a large number of documents with which to study any literature-based trends is very difficult outside of narrow domains, such as biomedicine (PubMed), physics (arXiv.org), and astronomy (Astronomical Data Service), where most of the literature is available via publicly accessible and machine-readable systems. In other domains, including those of interest within this project, the literature is spread across numerous publisher platforms that do not allow any overarching machine-accessibility capabilities. Google Scholar, while quite comprehensive in coverage, does not allow any significant automated literature mining. These methodological difficulties are a common refrain among numerous studies that have attempted to compile and analyze impact metrics for research infrastructures (Mayernik, et al, 2017). These scale-limiting copyright issues, along with journal editorial policies, publication platforms, and article formatting differences, present other uncontrollable factors for automating this kind of study.

EarthCollab - Linking via the Semantic Web

The third project of interest for this paper, called EarthCollab (active 2014-2018), has focused on using Semantic Web and linked data technology to facilitate the coordination and organization of complex scientific projects and their products. From a technology point of view, the goal of the project has been to develop information systems that demonstrate how the geosciences can leverage linked data to produce more coherent methods of information and data discovery for large multi-disciplinary projects and virtual organizations. The motivation for the project was to improve the discovery and sharing of information to advance research and scientific collaboration, enabling researchers to more easily find people, organizations, and research resources that are relevant to their work. The VIVO Semantic Web software suite (http://vivoweb.org/) was chosen because it is built around a web-centric data model that focuses on representing relationships between entities (Borner, et al., 2012). Figure 2 depicts how many-to-many relationships exist among (and within) scientific resources, projects, people, and organizations.

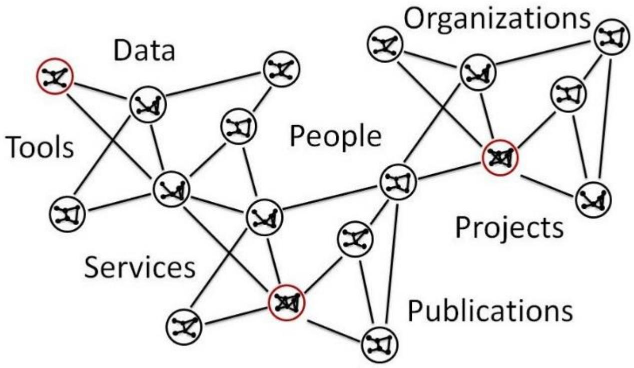

Figure 2. Networked science (Figure by Keith Maull)

The project used VIVO to provide web interfaces for researchers within two targeted scientific research areas to explore the people, publications, platforms, and data sets within their respective communities. These systems, called “Connect UNAVCO” (https://connect.unavco.org/) and “Arctic Data Connects” (http://vivo.eol.ucar.edu), have both been live for public use for over a year. As an illustration of the kinds of information that these systems contain, as of 1 April 2018, Connect UNAVCO contains records for nearly 6,000 scientific documents, 4,307 datasets, 3,841 research sites, 228 grants, 803 people, and 381 organizations. Arctic Data Connects, with a more focused case study on select arctic research projects, contains records for 354 datasets, 26 grants, 146 people, and 53 organizations. The information in these systems comes from a combination of existing metadata databases and newly created metadata. The resources are represented via the Semantic Web data model using multiple ontologies (Mayernik, et al., 2016). In the Semantic Web, ontologies are used to define entity types (classes), and the relationships between them (properties). Anything can be represented as a first-order object, as long as appropriate ontologies are declared. The result is a networked set of information, in which datasets, publications, scientific instruments, research projects, and research sites are each represented in the data model as distinct entities that have specific types of relationships with other entities. Practically, the VIVO software displays a web page for each entity, which provides information about the entity along with links to other entities that are related via explicitly declared relationships. For example, the web page for a data set within one of the systems will display all known links to associated publications, organizations, grants, creators, and instruments.

Another component of this project has been to undertake software development to add capabilities to the VIVO software suite. In the first two years of the project, we developed a prototype “cross-linking” approach to exchange information across VIVO instances. The motivation for this work was that participants in many scientific research projects are based at many different organizations. Some of those organizations already use VIVO to manage faculty and staff information profiles. The new cross-linking feature was developed to enable the exchange of information about specific people or entities that are in common across different VIVO instances. This new feature was intended to reduce duplication of information across systems and enable the distribution of authoritative information about a specific entity, such as when a single individual has VIVO profiles at different organizations. The cross-linking capability has been deployed within Connect UNAVCO and is being contributed to the core VIVO open source code base.

A key lesson learned within EarthCollab with regard to representing relationships between scholarly resources centers on the use and re-use of ontologies. To facilitate easier integration and data sharing across the geoscience community, our goal within this project has been to reuse existing ontologies as much as possible when developing project ontologies and web applications. Reusing existing ontologies did not prove to be as easy as originally anticipated. The difficult work in using ontologies is the conceptual modeling, in which the key entities and relationships of interest are specified and mapped. Only after this has been done can the next step take place, namely, finding relevant existing ontologies that map to your conceptual model, which is not a trivial exercise. In our project, no single ontology supported our project needs. We thus combined components from a couple of different ontologies and created custom ontology extensions as necessary to fill gaps (Mayernik, et al., 2016).

DISCUSSION

The three projects described above all focus on challenges associated with linking scholarly resources together in a web environment. The projects were conducted in parallel with some overlap in personnel, but no direct overlap in intended outcomes. The insights from the projects, however, are complementary. Together, they provide a larger view of the impediments and enablers of scholarly resource cross-linking efforts than any one of them could have allowed. This kind of “parallel trials” approach has proven to be particularly useful in situations where the optimal outcome (or path to an outcome) is not clear from the outset (Lenfle & Loch, 2010).

In this section, we provide an overarching discussion on scholarly resource cross-linking requirements, building on the knowledge and experience gained in working on multiple relevant projects concurrently. We first categorize the salient requirements and associated challenges in this problem space by type, and then depict how these requirements and challenges look different when looking forward in time vs. looking backward in time.

Requirements for Scholarly Resource Linking

A robust system for linking scholarly resources in a broad and sustainable fashion will have to navigate a set of complex and interrelated requirements. Just as research data and software creation, management, and preservation can be characterized through life cycle models (Carlson, 2014; Lenhardt, et al., 2014), we might characterize the requirements for a scholarly resource linking infrastructure through a "relationship life cycle" that encompasses the need to identify, validate, characterize, and preserve relationship and link information. Each of these areas is now discussed in turn.

Identifying relationships. Relationships between scholarly resources, to be useful for resource discovery, understanding, and use, have to be declared by somebody, or some entity. Where can (or should) these relationship declarations come from? Relationships that are not declared explicitly can sometimes be determined by computational processes, often relying on statistical measures to identify implicit relationships between specific entities or vocabulary terms included within a text (Sheth, Ramakrishnan, & Thomas, 2005). But these computational techniques are typically most successful in very specific applications and are difficult to generalize. The other obvious source for relationship information is from the creator(s) of the related resources. A truism of metadata generation for scholarly resources is that the scholars are best positioned to provide metadata about their own resources given their intimate knowledge of how those resources were produced and used. But efforts to gather information about relationships between data, software, publications, and other research resources directly from scholars also face well-known metadata creation challenges, namely that scholars have little incentive to describe such relationships explicitly, and that knowledge of specific relationships can decay quickly and be distributed among teams (Michener, et al., 1997; Edwards, et al., 2011). Gathering relationship information directly from scholars is thus a difficult task that does not scale well for the numbers of scholarly assets now being distributed on the web. Another potential source for relationship information is published literature. As we found in our Tracing Identifiers project, data and software citation are indeed increasingly cited formally, potentially making automated citation analysis possible. But, as our project also found, the timeline for these new citation practices to take root will likely be long. Identifying relationships is thus a significant rate-limiter for any effort to develop scholarly resource linking approaches. This challenge also has a bi-directional component, namely, even if relationship identification becomes more straight-forward, how do relationships propagate? Stated another way, if I know about a relationship between my resource and your resource, how do you learn about that relationship? How does my relationship declaration propagate to you? This requirement was what drove our interest in a relationship notification service within our Repository Cross-Linking project. There is currently no comprehensive aggregator for relationship information that could serve as this kind of general notification service. In the specific case of data-to-literature linking, the Scholix framework and associated Data-Literature Interlinking (DLI) Service have promise to serve key roles as a centralized third-party aggregator of relationship declarations (Burton, et al., 2017a; b), but these initiatives are still in early days of building adoption.

Validating relationships. Validation presents the next challenge and requirement. Whatever method used to identify relationships between scholarly resources, the relationship declarations need to be validated via some process, which again could be human or machine-based. A basic challenge for the validation process is simply to validate what entities are being related. Digital objects have diffuse boundaries and can change over time. Many approaches to relationship identification rely on the use of persistent identifiers (PIDs), such as DOIs, to ensure persistence and fixity. Technically, however, PIDs function as locators for the resources to which they are assigned, not as identifiers for those resources (Thompson, 2010; Duerr, et al., 2011). PIDs are also being assigned at different levels of granularity for different kinds of resources, and no broadly accepted rules exist to guide decisions about how PIDs should be assigned to compound digital objects (Mayernik, 2013). More practically, PID use is not comprehensive, as we studied in our Tracing Identifiers project. Many references to data, software, and other resources still take place through informal acknowledgements or in-text references. Validation of relationship declarations also depends on the notion of an authoritative source of the relationship information. Is relationship information that comes directly from resource creators, or from published articles, more authoritative than information from other sources? The DLI Service noted in the previous paragraph provides information about organizations that publish relationship declarations. But should all of providers be considered authoritative sources for relationship information? Mechanisms for ensuring trust in relationship declarations are key to any approach to relationship validation.

Characterizing relationships. For many purposes, the basic act of declaring that a relationship exists is not very useful. More information about the kind of relationships may need to be collected as well. The use of relationships, in other words, often requires additional description about the relationship itself. The Semantic Web approach is based on this premise, namely, that relationships should be designated as being of specific named types. Numerous ontologies and vocabularies have been created to define specific kinds of relationships that can hold between scholarly resources (see Mayernik, Phillips, & Nienhouse, 2016), but these relationship typologies are highly variable and inconsistent. In our EarthCollab project, we emphasized reusing existing ontologies as a way to represent entities and their relationships in an interoperable way. This proved to be more challenging than originally assumed due to the fact that numerous ontologies modeled the same kinds of resources (e.g. data sets) and relationships (e.g. citation relationships) in various ways (Mayernik, et al., 2016). An additional challenge in characterizing relationships is that different uses of relationship information might require different levels of description about the entities at either end of the relationships. The Scholix framework, for example, defines a very simple data model for representing relationships (Burton, et al., 2017b). For other applications, including our EarthCollab project, a simplistic data model did not support the goals of data discovery and understanding. A general issue related to characterizing relationships is that data models and metadata schemas do not always include ways to represent relationship information, or they have different requirements for describing relationships. The SHARE Notify data model, for example, did not have an explicit way to represent relationships during the time we were investigating its potential use as a relationship notification service within our Repository Cross-Linking project.

Preserving relationships. "Link rot" is the most obvious challenge for preserving relationships between web-based scholarly resources. Web sites change URLs or go down unpredictably, causing a cascade of errors for any links pointing toward the site that has gone down. This issue presents a significant challenge for projects that hope to use the scholarly literature as a source for relationship information. PIDs are again a solution for this issue, but PIDs are not inherently immune from link rot. All PID systems work through re-direct servers. Maintenance of redirects is an institutional challenge as much as it is a technical challenge. Organizations that register DOIs, for example, are required to maintain the resolution of their identifiers and associated landing pages. Beyond persistent resolution, preservation challenges relate to the level of description that might be needed to understand relationships as time goes by, and/or as the user communities for the resources and their relationships change over time. Processes to write and document the US National Climate Assessment report, for example, a cross-agency consensus report on climate change, have changed significantly in recent years to provide highly structured relationship information to illustrate specific linkages between scientific claims and the underlying data and research papers (Tilmes, et al., 2013). Baker, Duerr, & Parsons (2015) provide another vivid example of how user needs, and documentation needs, can shift significantly over time, even for the same scientific resource. It should be emphasized that in both of these examples, creating and curating this information was the responsibility of dedicated staff. Preserving the understandability of relationships is thus an ongoing process that can require significant expertise. As part of this preservation requirement, frameworks like Scholix and systems like the DLI Service should be explicit about their own sustainability models, to engender trust by potential adopters.

Looking Forward and Backward in Time

The issues associated with relationship identification, validation, characterization, and preservation present different challenges and requirements when looking forward vs. backward in time. The Repository Cross-Linking project workflows depicted in Figure 1 above provide a clear illustration of this. Looking forward in time, the challenges center on how to institute relationship declaration/identification as a routine and robust part of scholarly publishing, data and software archiving, and repository technologies. As Marchionini, et al., note, "[e]fficient capture of data including provenance and metadata is most easily done by working at the start of the process (2012, pg. 17)." In the scenario in which resources are being newly created and deposited into repository systems, it might be possible to gather relationship information as resources and relationships are produced.

Looking backward in time, linking scholarly resources involves mining available literature (and other available documentation), along with potentially querying scholarly resource creators directly about relationships that hold between existing resources. Literature mining can be manual or automated, as in our Tracing IDs project, with the attendant advantages and disadvantages of either approach. Copyright and licensing restrictions clearly limit current efforts toward literature mining, except in the academic specializations where open publishing models are well established. Directly contacting and working with resource creators to establish relationship information for existing resources has obvious limitations as well, and will likely only take place in the context of specific projects or applications where this information is desired. Within our EarthCollab project, for example, we are investigating protocols for querying researchers directly to gather information about linkages between data and scientific articles, because the articles themselves typically do not contain enough information for us to reliably assert the relationships ourselves.

Table 1 depicts the differing kinds of work required to support broad scholarly resource linking, when the focus is backward vs. forward looking. Each cell further breaks down the work requirements into "technical" and "institutional" work. As I discussed in a prior paper (Mayernik, 2016), scholarly resource curation encompasses far more than just technical work. Institutional factors like community norms and expectations, the availability of intermediaries to support curation work, standards development and adoption, and individual routines all play important roles in determining the success of curation efforts. Table

	Looking back in time	Looking forward in time
Relationship Identification	_Technical work_ – Data mining approaches for extracting PID-based references from published literature _Institutional work_ – Interfacing with publishers & funders to open scholarly literature to more comprehensive data mining	_Technical work_ – Development of relationship aggregators, and associated open web services _Institutional work_ – 1. Promoting consistent use of PIDs, 2. Developing and adopting community frameworks for relationship distribution (e.g. Scholix)
Relationship Validation	_Technical work_ – Data mining approaches for extracting informal references from published literature, with confidence estimates _Institutional work_ – Interfacing with publishers & funders	_Technical work_ – Developing/updating data models to support relationship source and trust assertions _Institutional work_ – Developing organizational trust networks for relationship declarations
Relationship Characterization	_Technical work_ – Fitting non-standardized relationship declarations into community ontologies and metadata schemas _Institutional work_ – Coupling entity and relationship descriptions to the needs of the target communities and/or applications	_Technical work_ – Developing/updating data models and metadata schemas to consistently represent relationships and the entities being linked _Institutional work_ – Coordinating agreements on relationship semantics within particular communities, or for particular applications
Relationship Preservation	_Technical work_ – Coupling link crawling and web archiving tools _Institutional work_ – Curating relationship information over time, iteratively updating relationships as necessary to support user needs	_Technical work_ – Developing/adopting packaging tools that ensure links between resources are not lost over time _Institutional work_ – 1. Developing sustainability approaches for relationship aggregators, 2. Curating relationship information over time

Table 1. Technical and institutional work needed to support scholarly resource linking.

1 thus outlines how technical and institutional developments are both critical to achieving robust infrastructures for scholarly resource linking. This table is meant to be illustrative, not exhaustive. Other writing cited in this paper present additional areas of emphasis for current and future work in this space.

CONCLUSION

Efforts to link scholarly resources in reliable and sustainable ways face numerous difficult challenges that span technical and institutional factors. This paper discusses three projects that all focused on one or more aspects of these challenges. It synthesized key outcomes and lessons that emerged from these three projects, and presents insights into four key requirement areas, namely, issues associated with the need to identify, validate, characterize, and preserve relationship information between scholarly resources. I also outlined how attempts to build out linking infrastructures going forward face different challenges than initiatives to gather and characterize links between already existing data, software, publications, and other scholarly resources.

By running multiple projects concurrently, we have taken a “parallel trials” approach to scope out what kinds of initiatives are more conducive to solving specific problems related to scholarly resource linking. Solving local requirements, e.g. the desire to link two specific local systems, can be approached with limited technical complexity. Developing cross-organizational and cross-institutional solutions to supporting resource linking, however, require standards and coordination work, and will necessarily have a longer timeline. No single technical system or infrastructure is likely to provide a general solution to the challenges identified in this paper due to the wide-ranging stakeholder communities involved. Iteration and institutional work will be key.

The information sciences have expertise and research capacity to be strong contributors to scholarly resource linking efforts going forward. The issues described in this paper touch on scholarly communication, metadata frameworks, bibliometrics, digital preservation, and web architectures, all of which are historically and currently areas of research within library and information science disciplines. Scholarly resource linking infrastructures will have to find ways to couple findings and tools associated with all of these areas to move forward productively.

ACKNOWLEDGMENTS

I wish to express my gratitude to the numerous collaborators at NCAR, Cornell University, and UNAVCO who have contributed to the three projects described above. The work described in this paper was supported by U.S. National Science Foundation (NSF) grants 1449668, 1448480, 1440293, 1440213, and 1440181. NCAR is funded by the NSF.

REFERENCES

Bean, C.A. & Green, R. (Eds). (2001). *Relationships in the Organization of Knowledge.* Boston, MA: Kluwer.

Baker, K.S., Duerr, R.E., & Parsons, M.A. (2015). Scientific knowledge mobilization: Co-evolution of data products and designated communities. *International Journal of Digital Curation*, 10(2): 110-135. http://doi.org/10.2218/ijdc.v10i2.346

Borgman, C.L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6): 1059-1078. https://doi.org/10.1002/asi.22634

Borner, K., Conlon, M., Corson-Rikert, J., & Ding, Y. (Eds.). (2012). *VIVO: A Semantic Approach to Scholarly Networking and Discovery.* San Rafael, CA: Morgan & Claypool.

Burton, A., et al. (2017a). The data-literature interlinking service: Towards a common infrastructure for sharing data-article links. *Program*, 51(1): 75-100. https://doi.org/10.1108/PROG-06-2016-0048

Burton, A., et al. (2017b). The Scholix Framework for interoperability in data-literature information exchange. *D-Lib Magazine*, 23(1/2). https://doi.org/10.1045/january2017-burton

Carlson, J. (2014). The use of life cycle models in developing and supporting data services. In J.M. Ray (Ed.), *Research Data Management: Practical Strategies for Information Professionals* (pp. 63-86). West Lafayette, IN: Purdue Univ. Press.

Duerr, R., Downs, R., Tilmes, C., Barkstrom, B., Lenhardt, W.C., Glassy, J., Bermudez, L., et al. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, 4(3): 1-22. https://doi.org/10.1007/s12145-011-0083-6

Edwards, P.N., Mayernik, M.S., Batcheller, A., Borgman, C.L., and Bowker, G.C. (2011). Science friction: Data, metadata, and collaboration in the interdisciplinary sciences. *Social Studies of Science*, 41(5): 667-690. https://doi.org/10.1177/0306312711413314

Fienberg, S.E., Martin, M.E., & Straf, M.L. (Eds). (1985). Part I: Report of the Committee on National Statistics. In *Sharing Research Data*. Washington, D.C.: National Academy Press. http://www.nap.edu/catalog/2033/sharing-research-data

Hitchcock, S. (2013). *The effect of open access and downloads ('hits') on citation impact: a bibliography of studies*. University of Southampton. https://eprints.soton.ac.uk/354006/

Kent, W. (1978). *Data and Reality: Basic Assumptions in Data Processing Reconsidered*. New York: North-Holland.

Kriesberg, A., Huller, K., Punzalan, R., & Parr, C. (2017). An analysis of federal policy on public access to scientific research data. *Data Science Journal*, 16: paper 27. https://doi.org/10.5334/dsj-2017-027

Lagoze, C.J. (2010). *Lost Identity: The Assimilation of Digital Libraries into the Web.* Ph.D. diss. Cornell University.

Lenfle, S. & Loch, C. (2010). Lost roots: How project management came to emphasize control over flexibility and novelty. *California Management Review*, 53(1): 32-55. https://doi.org/10.1525/cmr.2010.53.1.32

Lenhardt, W.C., Ahalt, S., Blanton, B., Christopherson, L., & Idaszak R. (2014). Data management lifecycle and software lifecycle management in the context of conducting science. *Journal of Open Research Software*, 2(1): e15. https://doi.org/10.5334/jors.ax

Marchionini, G., Lee, C.A., Bowden, H., & Lesk, M. (Eds.). (2012). *Curating for Quality: Ensuring Data Quality to Enable New Science*. Final Report: Invitational Workshop Sponsored by the National Science Foundation, Sept. 10-11, 2012, Arlington, VA. http://openscholar.mit.edu/sites/default/files/dept/files/altman2012-mitigating_threats_to_data_quality_throughout_the_curation_lifecycle.pdf

Mayernik, M.S. (2013). *Bridging data lifecycles: Tracking data use via data citations workshop report.* NCAR Technical Note, NCAR/TN-494+PROC, Boulder, CO: National Center for Atmospheric Research (NCAR). https://doi.org/10.5065/D6PZ56TX

Mayernik, M.S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, 67(4): 973-993. https://doi.org/10.1002/asi.23425

Mayernik, M.S. (2017). Open data: Accountability and transparency. *Big Data & Society*, 4(2). https://doi.org/10.1177/2053951717718853

Mayernik, M.S., Gross, M.B., Corson-Rikert, J., Daniels, M.D., Johns, E.M., Khan, H., … & Stott, D. (2016). Building geoscience Semantic Web applications using established ontologies. *Data Science Journal*, 15, article 11: 1-10. https://doi.org/10.5334/dsj-2016-011

Mayernik, M.S., Hart, D.L., Maull, K.E., & Weber, N.M. (2017). Assessing and tracing the outcomes and impact of research infrastructures. *Journal of the Association for Information Science and Technology*, 68(6): 1341-1359. https://doi.org/10.1002/asi.23721

Mayernik, M.S. & Maull, K.E. (2017). Assessing the uptake of persistent identifiers by research infrastructure users. *PLoS ONE*, 12(4): e0175418. https://doi.org/10.1371/journal.pone.0175418

Mayernik, M.S., Phillips, J., & Nienhouse, E. (2016). Linking publications and data: Challenges, trends, and opportunities. *D-Lib Magazine,* 22(5/6). https://doi.org/10.1045/may2016-mayernik

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., and Stafford, S.G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1): 330-342.

Mitroff, I.I. (1974). Norms and counter-norms in a select group of the Apollo moon scientists: A case study of the ambivalence of scientists. *American Sociological Review*, 39(4): 579-595. https://doi.org/10.2307/2094423.

Pepe, A., Mayernik, M., Borgman, C.L., & Van de Sompel, H. (2010). From artifacts to aggregations: Modeling scientific life cycles on the semantic web**.** *Journal of the American Society for Information Science and Technology*, 63(3): 567-582. https://doi.org/10.1002/asi.21263

Piwowar, H.A. & Vision, T.J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1: e175. https://doi.org/10.7717/peerj.175

Sheth, A., Ramakrishnan, C., & Thomas, C. (2005). Semantics for the Semantic Web: The implicit, the formal and the powerful. *International Journal on Semantic Web and Information Systems*, 1(1): 1-18. https://doi.org/10.4018/jswis.2005010101

Silvello, G. (2017). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, 69(1): 6-20. https://doi.org/10.1002/asi.23917

Thompson, H.S. (2010). What is a URI and why does it matter? *Ariadne*, 65. http://www.ariadne.ac.uk/issue65/thompson-hs/

Tilmes, C., Fox, P., Ma, X., McGuinness, D.L., Privette, A.P., Smith, A., … Zheng, J.G. (2013). Provenance representation for the National Climate Assessment in the Global Change Information System. *IEEE Transactions on Geoscience and Remote Sensing*, 51(11): 5160-5168. https://doi.org/10.1109/tgrs.2013.2262179

Van de Sompel, H. & Nelson, M. L. (2015). Reminiscing about 15 years of interoperability efforts. *D-Lib Magazine*, 21(11/12). http://doi.org/10.1045/november2015-vandesompel

Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., & Warner, S. (2004). Rethinking scholarly communication. *D-Lib Magazine*, 10(9). https://doi.org/10.1045/september2004-vandesompel

Willinsky, J. (2005). The unacknowledged convergence of open source, open access, and open science. *First Monday*, 10(8). http://ojphi.org/ojs/index.php/fm/article/view/1265/1185

Woelfle, M., Olliaro, P., & Todd, M.H. (2011). Open science is a research accelerator. *Nature Chemistry*, 3(10): 745-748. http://doi.org/10.1038/nchem.1149