2015

Reminiscing About 15 Years of Interoperability Efforts

Herbert Van de Sompel
Los Alamos National Laboratory

Michael L. Nelson
Old Dominion University

Original Publication Citation

De Sompel, H.V., & Nelson, M.L. (2015). Reminiscing about 15 years of interoperability efforts. *D-Lib Magazine, 21*(11-12), 1-12. doi: 10.1045/november2015-vandesompel

D-Lib Magazine

November/December 2015
Volume 21, Number 11/12
Table of Contents

Reminiscing About 15 Years of Interoperability Efforts

Herbert Van de Sompel
Los Alamos National Laboratory
herbertv@lanl.gov

Michael L. Nelson
Old Dominion University
mln@cs.odu.edu

DOI: 10.1045/november2015-vandesompel

Printer-friendly Version

(This Opinion piece presents the opinions of the authors. It does not necessarily reflect the views of D-Lib Magazine, its publisher, the Corporation for National Research Initiatives, or the D-Lib Alliance.)

Abstract

Over the past fifteen years, our perspective on tackling information interoperability problems for web-based scholarship has evolved significantly. In this opinion piece, we look back at three efforts that we have been involved in that aptly illustrate this evolution: OAI-PMH, OAI-ORE, and Memento. Understanding that no interoperability specification is neutral, we attempt to characterize the perspectives and technical toolkits that provided the basis for these endeavors. With that regard, we consider repository-centric and web-centric interoperability perspectives, and the use of a Linked Data or a REST/HATEAOS technology stack, respectively. We also lament the lack of interoperability across nodes that play a role in web-based scholarship, but end on a constructive note with some ideas regarding a possible path forward.

1 Introduction

We find ourselves two decades into the transition of research communication from a paper-based endeavor to a web-based digital enterprise, and well into the transition of the research process itself from a largely hidden activity to one that becomes plainly visible on the global network.

By nature, scholarship and scholarly communication are highly distributed activities. Despite a series of acquisitions and mergers within the scientific publishing industry, it is not a "winner takes all" environment but rather one that consists of an ever increasing number of nodes on the network that play a part. It is not an environment in which a single player can dictate others to act as it

wishes them to. Hence, in order to migrate this distributed activity from a gathering of silo-ed nodes to an ecology of collaborating nodes, establishing cross-node interoperability is essential.

Considering that many thousands of nodes are at play, achieving a significant level of interoperation is far from trivial. And it definitely has not been achieved. Most scholarly nodes can best be characterized as stand-alone portals, destinations on the web, rather than infrastructural buildings blocks in a global, networked scholarly communication system. The exception to this consideration is the interoperation that has been achieved with regard to identification, through the broad adoption of DOIs — introduced about 20 years ago — and more recently ORCID — a crucial component, but, conceptually a close cousin of DOI.

It is fair to say that interoperation has not been a focus for the scholarly community. One can only hope that this will eventually change, in order to maximize the significant investments in research infrastructures made by funding agencies, and to achieve frictionless research communication and collaboration. But, interoperability can be achieved in different ways and some approaches may have more potential than others. Having been involved in several interoperability efforts during the past 15 years, we felt that it might be informative to describe the evolution of our thinking with that regard.

In the remainder of this paper, we zoom in on OAI-PMH (1999), OAI-ORE (2006), and Memento (2009) and reflect on the world-view and technical foundations upon which they were built. We draw some conclusions and propose to Signpost the Scholarly Web as a means to achieve broad, coarse, yet meaningful, interoperation across nodes in the web-based scholarly communication and research environment.

Readers will note that links to web resources in this article have been "decorated" with actionable attributes as per the Robust Links specification. The down arrows next to each link display a menu with robust links to archived snapshots of the referenced resources. These robust links can be followed if the original link no longer works or if it is desirable to see what the linked content was around the time the original link was put in place. This Robust Link approach is aimed at avoiding link rot and content drift, and results from the Mellon-funded Hiberlink project. It leverages the Memento protocol, discussed in this paper, and associated infrastructure.

2 The Open Archives Initiative Protocol for Metadata Harvesting — OAI-PMH (1999)

The Open Archives Initiative (OAI) started in 1999 as an heroic effort to transform scholarly communication. In the invitation that was sent out for the Santa Fe meeting that led to its creation, the effort's initial goal was stated to be working towards the creation of a universal service for scholarly e-prints, a term used to include both unreviewed preprints and openly available peer-reviewed papers. The universal service was thought of as a fundamental and free layer of scholarly information, above which both free and commercial services could flourish.

The OAI was a technical initiative and hence focused on matters of interoperability as a way to break ground for a universal adoption of e-print communication mechanisms. The initial interoperability goal was to improve the discoverability of e-prints, actually making them easier to discover than journal publications hosted on publisher portals. The result of this focus was OAI-PMH, a protocol for the recurrent exchange of metadata between data providers (e-print repositories) and service providers (Figure 1), which was to an extent inspired by the Dienst protocol.

Figure 1: Diagram Used to Illustrate OAI-PMH (2000)

In hindsight, the choice to support discovery by means of metadata exchange may be surprising. Yet, in those days, many repositories only held metadata [12]. And those that hosted full text made it available in PDF or PostScript, formats that were not yet indexed by web search engines. Furthermore, downloading and storing full-text in 1999 was an expensive proposition! It may be equally surprising that the discovery problem was not approached from the web-centric perspective of search engine optimization. Yet, although Google existed, it was not the hegemonic force that it is today. And, other search engines such as AltaVista and Lycos didn't truly impress. Generally speaking, much of the repository materials that had to be surfaced were in the deep web, not accessible to standard web crawlers [6,8]. Furthermore, until the mid-2000s, most search engines did not crawl entire web sites; they often stayed at the "top" of sites and only lightly sampled links that pointed deeper into a site [11]. Both realities, the deep web and incomplete crawling, motivated the creation of a special-purpose protocol that could guarantee full coverage of a repository. We briefly considered the RSS syndication formats (Atom did not yet exist) as the basis of desired functionality. But, in 1999, these formats were in flux and had limited expressiveness. For example, none had the ability to handle multiple metadata formats simultaneously. Hence, a repository add-on was defined to expose metadata to parties that would hopefully use it to create attractive discovery services.

An inspection of some design choices made in the protocol reveals a lack of understanding of the very essence of the web and its primitives, namely resources, URIs, HTTP, links, media types and content negotiation:

- In order to obtain a metadata record, a `GetRecord` verb and a record identifier are transported as query components on an HTTP GET request. In a web-centric approach, the HTTP URI would be the identifier of the metadata record and one would issue a GET on it. The record's desired metadata format is also transported as a query component, whereas this preference could be expressed by means of a media type in an `Accept` HTTP request header used in content negotiation. In addition, the metadata record is delivered in an OAI-PMH wrapper, making it impossible to use the response without unpacking it. More generally, the use of the section title "HTTP Embedding of OAI-PMH requests" in the protocol specification indicates that a protocol was specified and that this protocol was then tunneled over HTTP. The protocol was not defined in terms of available HTTP functionality.
- Paging among incomplete responses to `ListRecords` and `ListIdentifiers` requests is achieved by means of a `resumptionToken` handed out by the server and returned in the subsequent request by the client. While such an approach was common in pre-web protocols such as Z39.50, a web-centric approach would implement paging using fully constructed links without requiring client-side knowledge regarding the use of the `resumptionToken` to construct the link to additional content.

OAI-PMH has been very broadly adopted even beyond the target scholarly communication environment. Its initial success was largely related to its close association with the early Open Access movement for which it was an important catalyst. Despite its wide adoption, the protocol has been rightly criticized for choices like the aforementioned ones. Let it be clear that those were not made with ill intentions nor out of ignorance on the part of those involved in the design, testing, and implementation of OAI-PMH. Rather, they are signs of the times. Let's not forget that Roy Fielding's REST thesis [4] was only published in 2000, and that the W3C's Architecture of the World Wide Web was drafted in 2002 and finalized in 2004. The implication of these seminal

documents could have been deduced from the early HTTP RFCs (RFC2616, RFC2068, RFC1945). But, like many others, we had not grasped it. The very existence of technologies such as SOAP (ca. 2000) nicely illustrates the prevailing mindset of RPC-style design and of HTTP as a replaceable transport. The latter shouldn't come as a surprise because gopher had just been thrown overboard, and it hadn't been that long since anonymous FTP had been the primary method of resource discovery and transfer. This lead to a perspective that transport protocols like HTTP were transitory and application semantics needed to be defined independently. Also, SQL databases were omnipresent and the focus was on exporting their content to the web. The database script that was able to achieve this was the center of attention and HTTP was merely seen as a transport protocol to make that happen. As a result, in many cases, ad-hoc protocols were tunneled over HTTP. Similarly, the database records were considered important, not their web equivalent resources.

Technically, OAI-PMH would now be referred to as RESTless or unRESTful. Conceptually, we have come to see it as repository-centric instead of resource-centric or web-centric. It has its starting point in the repository, which is considered to be the center of the universe. Interoperability is framed in terms of the repository, rather than in terms of the web and its primitives. It is a perspective in which a repository resembles a brick and mortar library, a library that one can go visit and that allows such visits subject to policies — the protocol — that may simultaneously be well intended and idiosyncratic. This kind of repository, although it resides on the web, hinders seamless access to its content because it does not fully embrace the ways of the web.

---

3 The Open Archives Initiative Object Reuse and Exchange — OAI-ORE (2006)

The 2004 paper "Rethinking Scholarly Communication" [13] described how the digital environment allows to fundamentally reimagine scholarly communication as an ecosystem of distributed and interoperating service nodes, each fulfilling one of the core functions — registration, certification, awareness, archiving, rewarding [10]. By observing the changing nature of scholarly research, the paper noted that the units of communication that would flow through a future scholarly communication system would no longer be atomic papers or monographs. Rather, they would be compound objects, consisting of multiple resources connected by various relationships and interdependencies.

These observations led the Open Archives Initiative to embark on a new endeavor, Object Reuse & Exchange. The effort was generally aimed at developing specifications to allow distributed repositories to exchange information about their constituent digital objects but eventually focused on modeling the compound objects to be used in the emerging scholarly communication system.

By 2006, nothing stood in the way anymore of obtaining a profound understanding of the core technologies that made the web tick. Nevertheless, the initial thinking for OAI-ORE followed in the footsteps of OAI-PMH. This can clearly be seen in the grant proposal that led to OAI-ORE. It never mentions HTTP as a fundamental enabling technology, and states, for example, that the specifications that will result from the effort "will describe common data models and interfaces for exchange of information based on these data models" thereby overlooking that HTTP's uniform interface provided all required methods (GET, POST, PUT, DELETE). This initial repository-centric perspective is also apparent in early architectural drawings emerging from the effort (see Figure 2 left) that come across as OAI-PMH on steroids. This line of thinking can, to an extent, be explained by the strong pull of repository-centric thinking and the background in digital library practice. Also, attempts [2,14] had already been made to use OAI-PMH to not only deal with metadata but also with actual content, typically full-text articles. These efforts resorted to the use of XML-based complex object formats such as METS or MPEG-21 DIDL, which actually could be transported using OAI-PMH and which contained content either by-reference or by-value.

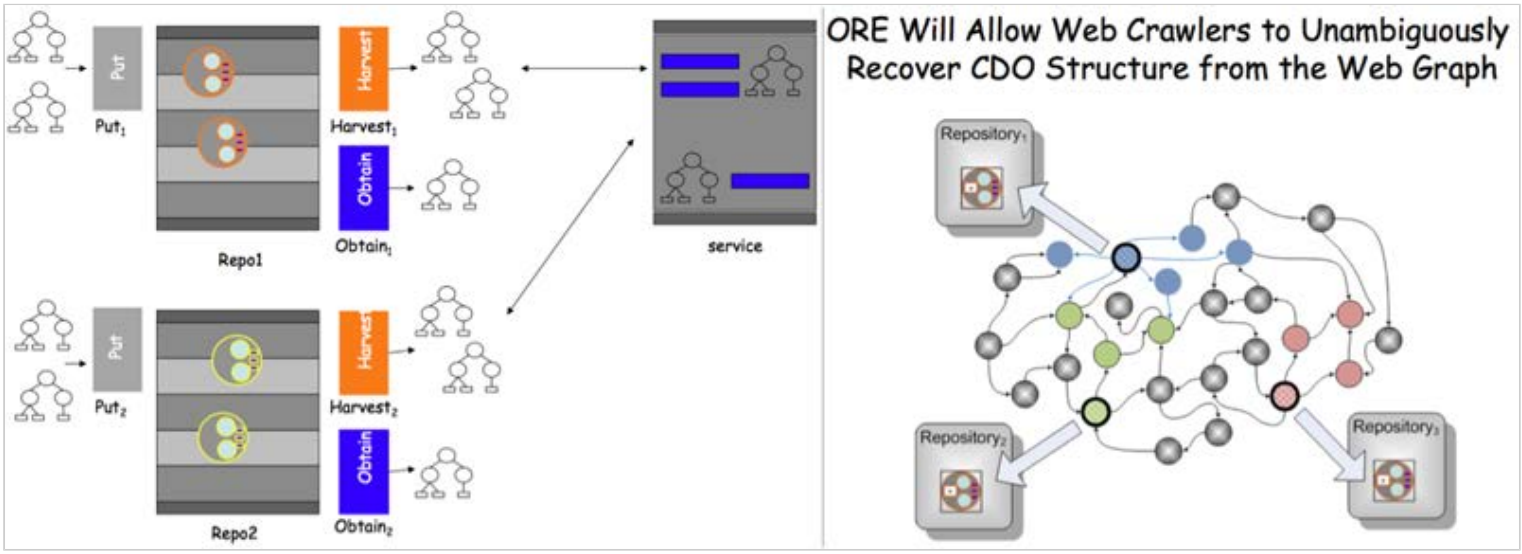

Figure 2: Diagrams Used to Describe OAI-ORE in August 2006 (left) and April 2007 (right)

However, a few months into the project, the perspective fundamentally changed to become web-centric, as illustrated in Figure 2 right, in which CDO stands for Compound Digital Object. Compound objects are no longer considered from the perspective of repositories but rather from the web. When looking from the web at a repository, one actually does not see a repository. Rather, one sees web resources that happen to be exposed by the repository. The notion of a repository does not exist in the Architecture of the World Wide Web; it does not even entail the notion of a web server. The Web is all about resources identified by URIs. In this web-centric perspective, a compound object consists of any number of URI-identified resources that exist somewhere on the web. Constituent resources may be hosted by one or more repositories but that is not considered an essential characteristic. This 180 degree shift in perspective became a veritable aha moment for us, a Kool-Aid we drank, something we saw that can never be unseen.

This shift turned the challenge for the OAI-ORE effort into providing a way to draw a line around the resources that were part of a specific compound object and to provide a URI identity to the union of those resources. We understood that the former could be achieved by publishing a document to the web listing the URIs that are part of the compound object. The approach to tackle the latter became clear when discovering the seminal paper about Named Graphs [3], which itself led to a Linked Data Tutorial and a collaboration with one of the movement's visionaries, Chris Bizer. The path towards a solution to deal with compound objects, meanwhile reframed as resource aggregations, based on HTTP and RDF had been paved.

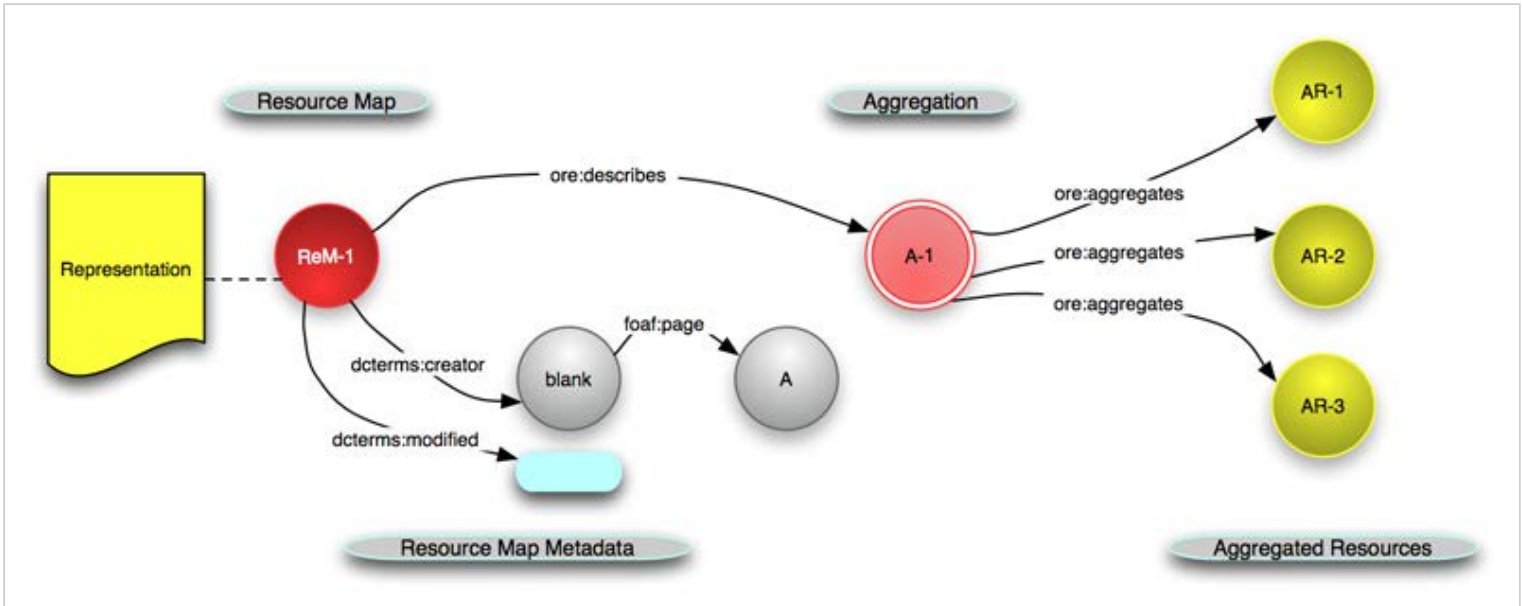

Figure 3: Architectural Diagram for OAI-ORE Showing Resource and Relation Types (2008)

In essence, the OAI-ORE specification prescribes publishing a machine-actionable document on the web that describes and provides identity to an aggregation of web resources. The relation types and properties introduced to achieve this are defined by means of an RDFS/OWL vocabulary (Figure 3). OAI-ORE also allows expressing relationships and properties pertaining to the aggregation, its description, and its aggregated resources. It also provides mechanisms to discover such a description from the HTTP URI that identifies the aggregation following the guidelines provided by Cool URIs for the Semantic Web. Anyhow, OAI-ORE also provides several ways to serialize the descriptions of aggregations, including RDF/XML, RDFa, and Atom; recently JSON-LD was added. The Atom serialization is a veritable disaster, and should be discontinued. It was created in response to community pressure because Atom was a rather popular technology at the time of the effort. But mapping the RDF-based model to the Atom model, and transporting RDF/XML in an Atom feed, led to a serialization that is both complicated and inelegant.

The Cool URI for the Semantic Web guidelines have been the topic of the heated HTTPRange14 debate, related to the use of HTTP URIs to not only identify documents but also real-world things. The need for distinct URIs to, respectively, identify and describe the aggregation also caused quite some confusion and discussion in the course of the OAI-ORE effort, which, to many involved, was the first exposure to Linked Data concepts. We can't keep ourselves from remembering that some Linked Data people were pretty glib about creating an abundance of URIs as if they were going to be free forever. As such, we observe with some amusement that the URI where an initial version of the Cool URIs spec was available, http://www.dfki.uni-kl.de/~sauermann/2006/11/cooluris/, is meanwhile 404. A Memento is at https://archive.is/20070307124458/http://www.dfki.uni-kl.de/~sauermann/2006/11/cooluris/.

In the course of the effort, the Prototyping Team at LANL conducted an experiment related to archiving compound objects that convincingly illustrated the advantages of the web-centric approach; a screencam documenting the experiment is still available. The starting point of the experiment was the consideration that future units of scholarly communication would not only be compound but also dynamic. That is, constituents resources of an aggregation could change over time and the aggregation itself could change, with constituents resources entering and leaving an aggregation. Since the aggregation approach used in OAI-ORE is by-reference — a description refers to constituents by means of their HTTP URI instead of embedding them — archiving an aggregation requires visiting and revisiting both the evolving description and the constituent resources. Since the descriptions are web documents with embedded URIs, this requirement seemed quite similar to web crawling and web archiving of HTML documents. Hence, the team used off-the-shelf tools, Heritrix to crawl and capture, and Wayback to store and make accessible the various states of evolving aggregations. We will never know what a repository-centric solution to the compound object problem would have looked like, but chances are significant that tackling this very archival challenge would have required special-purpose tools.

OAI-ORE support has been implemented in various efforts in the realm of scholarly communication, cultural heritage, and learning. It forms the core of the Research Object model [1] that, in addition to aggregation and identification, supports annotation and provenance.

---

4 Memento — Time Travel for the Web

The creation of the Memento protocol was directly related to an obsession with the temporal dynamics of the web, as exemplified by the aforementioned experiment, and by an ongoing interest in web archiving. Thanks to web archives, it had become possible to revisit traces of the past of the web, by visiting a web archive's portal and looking up an original URI. We observed that public web archives were becoming increasingly common and discussions about web archiving were no longer limited to the Internet Archive's Wayback Machine. WebCite, a pioneer in combating reference rot, had been in operation for a while and Archive-It, although part of the Internet Archive, had a distinct collection and access mechanism. But web archives remained silo-ed destinations on the web, with archival resources decoupled from their originals on the live web. Lacking was an HTTP-level solution to use an original URI to revisit a prior representation as it existed at a specific date/time, or, to use the URI of such a prior version to access the current representation. A similar consideration applied to resource versioning systems: specifications existed that detailed interoperability for navigation between versions (RFC5829, RFC5988), but protocol-level access to the version that existed at a specified date/time did not exist. In 1996, Tim Berners-Lee had hinted at the need for such a capability when discussing the genericity of resources. But the HTTP specification was released without a solution to deal with time, although it offered content negotiation as a way to address two other dimensions of genericity: language and content-type.

Memento, specified in RFC7089, introduces interoperability for time-based access to resources' versions. Using a Memento client like Memento for Chrome one can literally set a browser to some date in the past and navigate the web as it was then, visiting resource versions that reside in distributed web archives and versioning systems. This time travel capability is achieved in a rather straightforward manner, merely using the primitives of the web (see Figure 4). The Memento protocol introduces datetime negotiation, a variant of content negotiation. And it introduces a TimeGate, a resource associated with an original resource that is

capable of providing access to the resource version that is temporally closest to a preferred datetime expressed by the client. In addition, the protocol uses typed links — conveyed in the HTTP Link header as per Web Linking (RFC5988) — to point, among others, from an original resource to an associated TimeGate and from a version resource — meanwhile commonly referred to as a Memento — to the original resource. These relation types are registered in the IANA Link Relation Type Registry. In addition, the protocol uses some existing HTTP headers "as is", introduces new content for some others, and introduces some new headers. The latter are registered in the IANA Message Headers Registry. Memento is a textbook example of the architectural style known as HATEOAS, Hypertext As the Engine Of Application State, a constraint on the REST application architecture.

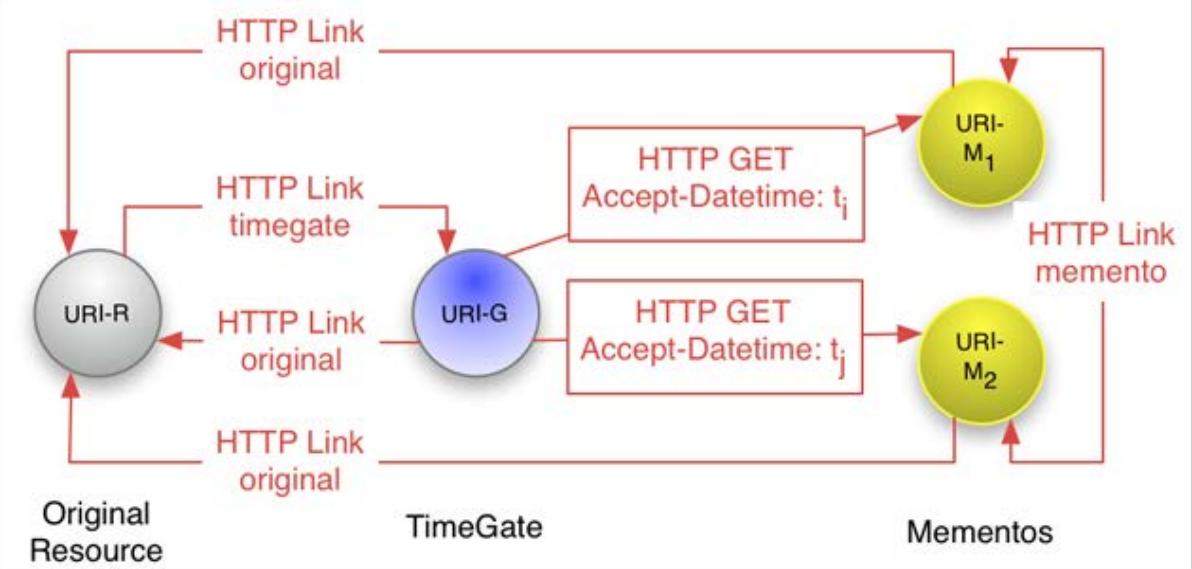

Figure 4: Architectural Diagram for Memento showing Datetime Negotiation and Typed Links (2010)

The Memento protocol solves a fundamental problem related to web archives that had been pointed out by Masanes [7]: web archives were destinations on the live web that lacked any form of interoperability. Memento turned web archives into archival infrastructure, tightly interconnected with the live web. When announced at the end of 2009, it generated a storm of positive reactions in the press, worldwide. But it was also the subject of criticism from privacy advocates who loathed Memento's power to make discovering previously published web resources easier. We submitted a paper detailing aspects of the Memento solution to WWW2010 [15] but were dismissed by a reviewer who wondered: "Is there any statistics to show that many or a good number of Web users would like to get obsolete data or resources?" [sic]. We found comfort in the knowledge that Tim Berners-Lee's paper about the World Wide Web had been rejected from the Hypertext conference in 1992. Memento is meanwhile supported by most public web archives, including the massive Internet Archive. Adoption by version control systems has been slow but our hopes remain high that, eventually, major portals such as Wikipedia and GitHub will add support.

5 Discussion

Both OAI-ORE and Memento take a web-centric, resource-centric approach towards interoperability, in contrast to the repository-centric approach taken by OAI-PMH. Their tools of the interoperability trade are:

- The basic ingredients of the web, as described in the Architecture of the World Wide Web: Resource, URI, HTTP as the uniform interface for resources (HEAD/GET, POST, PUT, DELETE), Representation, Media Type, Link, and Content Negotiation.
- Both OAI-ORE and Memento use content negotiation:
  - OAI-ORE uses content negotiation to distinguish between URIs that identify real-world objects and URIs of information resources that describe them (as per Cool URIs for the Semantic Web), and also to request a preferred serialization of the description of a resource aggregation.
  - Memento negotiates in the time dimension to request a resource version as it existed at a preferred date/time.
- Both OAI-ORE and Memento make use of typed links, but there are significant differences:
  - OAI-ORE uses the Linked Data approach based on RDF/RDFS/OWL to define and convey typed links contained in machine-actionable documents that are published to the web. A prime goal of the typed links is expressing that

resources adhere to the data model that OAI-ORE introduced to handle resource aggregations.

- Memento uses the REST/HATEAOS approach with a choice for the Web Linking technique whereby links with IANA-registered relation types are conveyed by means of HTTP Link headers. The prime goal of the typed links is to allow a client to navigate from one resource to another, to transition from one application state to another.

The Linked Data approach used by OAI-ORE is also the basis of several other scholarly communication interoperability efforts, including Open Annotation, PROV, Research Objects, and SharedCanvas. The approach is appropriate to achieve interoperability when data models aimed at expressing certain resource types is concerned [9]. It can be used for broad, cross-community models (e.g. the OAI-ORE model for resource aggregation), as well as for focused, community-specific models (e.g. the SharedCanvas model for describing digital facsimiles of physical objects). It comes with a non-trivial barrier to entry related to the use of technologies that, even today, remain unfamiliar to many players in the scholarly communication environment.

The REST/HATEAOS approach used by Memento, is also at the core of ResourceSync, which should be considered the next-generation OAI-PMH. ResourceSync is based on the Sitemap protocol that is widely used for search engine optimization; typed links are mainly conveyed in XML documents, not in HTTP headers. The approach has also been explored in the context of interoperability for scientific computation in environmental modeling [5]. Numerous RESTful APIs have been introduced but many fail to adhere to the HATEAOS constraint. Generally speaking, despite its low barrier to entry, a direct consequence of the use of the omnipresent HTTP stack, the REST/HATEAOS approach remains under-explored. Nevertheless, it is very well-suited to empower a client to navigate the web-based scholarly environment in an informed manner, and to allow it to makes choices among various possible transitions from its current state to a new one. As a matter of fact, the combination of HTTP's uniform interface, typed links, and media types forms an API that a client can utilize to achieve its application goal [9].

---

6 A Way Forward — Signposting the Scholarly Web

Most nodes that play roles in web-based scholarly communication and research currently focus on catering to human users. And, many of those that do cater to machines take a repository-centric stance towards interoperability. They expect a client to interoperate with them — a veritable challenge if a client needs to deal with many nodes in the scholarly web — instead of them aligning with the ways of the web. While some nodes do use current technologies to cater to machines, others, for example institutional repositories, are still stuck in the status quo of the late nineties.

It is high time to start working towards increased levels of web-centric interoperability across nodes. The interoperability infrastructure that has been introduced with regard to identification, particularly CrossRef DOIs, convincingly illustrates that catering to machines leads to improved services for both humans and machines. But, it is high time to move beyond identification only. Such a statement, coming from interoperability geeks like ourselves, may not carry enough convincing power, especially since we have repeated it uncountable times in the past years. So, we refer to an excerpt of a recent blog post by Cameron Neylon to get the job done:

> *"Infrastructures for identification, storage, metadata and relationships enable scholarship. We need to extend the base platform of identifiers into those new spaces, beyond identification to include storage and references. If we can harness the benefits on the same scale that have arisen from the provision of identifiers like Crossref DOIs then the building of new services that are specific to given disciplines will become so much easier. In particular, we need to address the gap in providing a way to describe relationships between objects and resources in general. This base layer may be "boring" and it may be invisible to the view of most researchers. But that's the way it should be. That's what makes it infrastructure."*

If we were "King of Scholarly Communication", we would issue a decree mandating the REST/HATEAOS path using IANA-registered link relation types to establish coarse-grained interoperability among scholarly communication and research nodes. That path relies on a basic and widely used technological stack, which should make the threshold for broad adoption across the significant amount of nodes surmountable. This path does not stand in the way of achieving interoperability based on the Linked Data path. But, in our opinion, because of its reliance on a more advanced technology stack, the latter path stands more chance of being successful when trying to achieve a fine-grained, deep, level of interoperability among a select set of nodes rather than across the entire web-based scholarly environment.

Under the umbrella "Signposting the Scholarly Web", we have started to explore the REST/HATEAOS path. Our approach thus far has been to select patterns that commonly occur in web-based scholarship. They are patterns that are easily discernible by humans but impossible to grasp by machines. Hence, we have investigated how these patterns could be revealed to machine agents by using Web Linking and links with IANA-registered relation types. These are examples of patterns we looked into:

- The landing page pattern (see Figure5, top) — Scholarly assets are commonly identified by persistent identifiers, e.g. a `DOI`. In order to access the identified resource, the identifier is put on an HTTP URI, e.g. `http://dx.doi.org/<DOI>`. Dereferencing that HTTP URI results in a redirect to a landing page, which contains links to resources that are constituents of the identified assets and others that are not. While humans can easily grasp this pattern, a machine sees a bunch of HTTP URIs and links without understanding their relationship. Recently, we pointed at the urgent need to clarify this omnipresent pattern [16]. In a presentation for the FORCE2015 conference, for which a video recording is available, we showed how this pattern can be clarified using the REST/HATEAOS approach. More recently, we have explored it in some depth with representatives from CrossRef and arXiv.org. Tackling this patterns allows a machine agent, among others:
  - To understand that the splash page describes the DOI-identified asset;
  - To determine that resource A is not part of the DOI-identified asset;
  - To navigate towards the profile of the authors of the asset when landing on any of the constituent resources of the DOI-identified asset;
  - To understand that a DOI is associated with the PDF, HTML, and JPEG resources and that this DOI should preferably be used to refer to those resources;
  - To associate annotations made to the HTML page with the DOI.

  Note that this pattern can also be addressed using OAI-ORE, but, to the best of our knowledge, this has not happened.

- The resource version pattern — Increasingly, scholarly assets are versioned. We have described in detail how to express temporal resource versioning in a manner that aligns with the Memento protocol, yet does not necessarily require full compliance with it. Adopting the described pattern allows a machine client, among others:
  - To navigate between versions;
  - To obtain a standardized list of versions and their creation date/times;
  - To understand what the generic URI is from which, at any moment in time, the then-current resource version is available;
  - To access the version that was operational at a specified date/time.
- The snapshot pattern (see Figure 5, bottom) — Researchers commonly develop code in GitHub. And, driven by a desire to reference that code in a manner that has become the norm in scholarship, they deposit a snapshot of the code in FigShare or Zenodo. In return they receive a DOI and a promise that the snapshot will be archived. Automated mechanisms have been put in place to streamline that process. The snapshots reveal their provenance in GitHub to human visitors but machines are left in the dark as to the intricate relationship between the snapshot and the original GitHub repository. This pattern can be clarified using links with IANA-registered relation types and by doing so allows a client, among others:
  - To understand that the snapshot is a duplicate of a specific version of the code in GitHub;
  - To navigate from the snapshot to the current state of the GitHub repository;
  - To navigate from the snapshot to the version of the code in GitHub that was operational at a specified date/time, if, eventually, GitHub supports the Memento protocol (see the red links in Figure 5);
  - To navigate from the specific version of the code in GitHub to its snapshot;
  - To understand that a DOI is associated with a specific version of GitHub code.

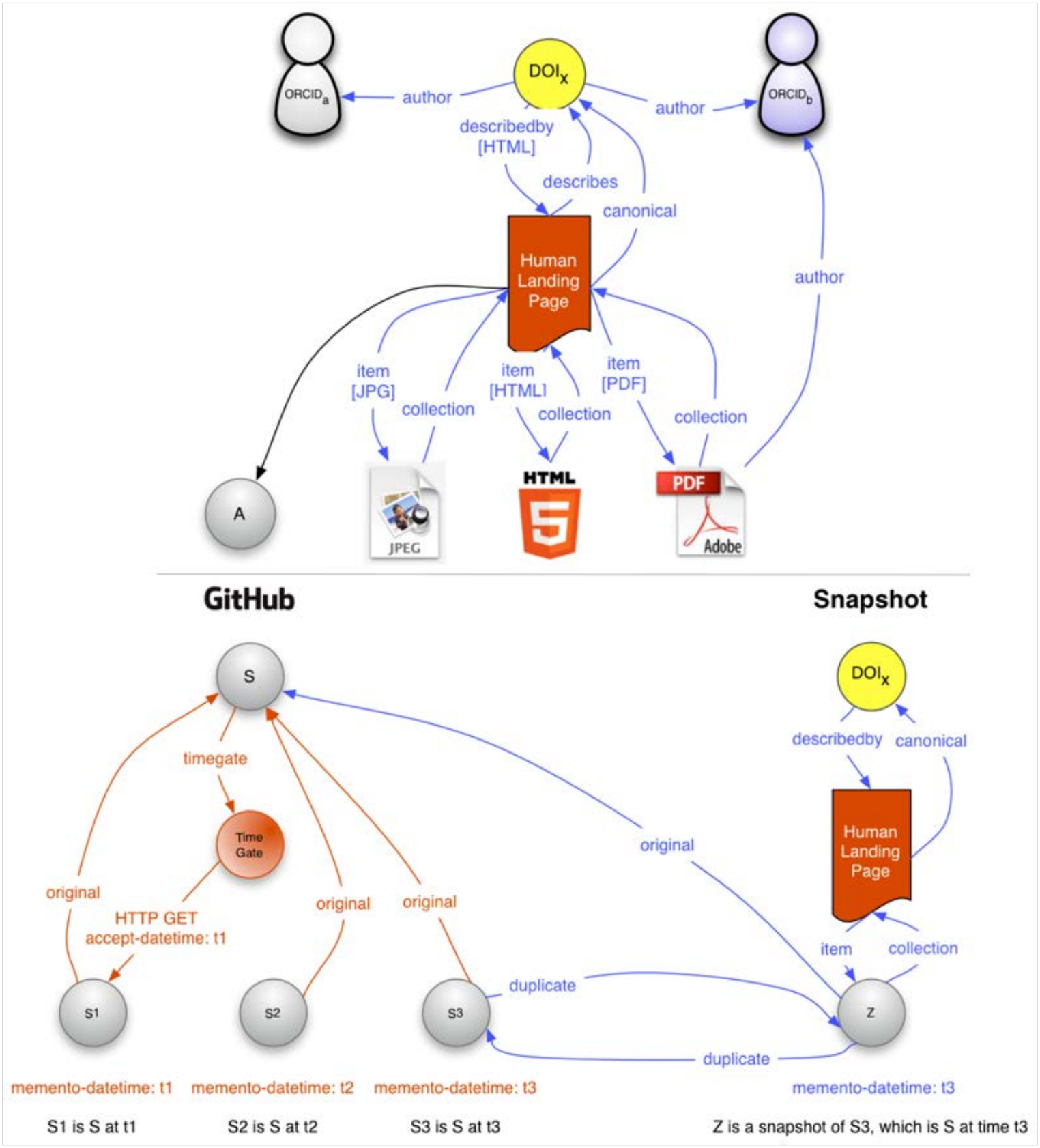

Figure 5: Typed Links Used to Clarify the Landing Page Pattern (top) and the Snapshot Pattern (bottom)

7 Conclusion

We have been involved in interoperability efforts related to web-based scholarship for over 15 years. We have gained a web-centric

perspective towards interoperability and are convinced it is the only viable and sustainable path forward. Interestingly, the REST/HATEOAS approach that we promote is less widely adopted on the web at large as one might expect. Rather frequently, web-based companies create custom protocols — said to be RESTful — that share a serialization format (JSON) but not semantics. In doing so, they behave in a manner that more closely resembles repository-centric than web-centric thinking, despite the use of contemporary technologies. We think that this approach is closely related to the desire of these companies to achieve monopoly positions. In such a world-view, there is less need or pressure to truly interoperate with parties other than the customers. However, such a perspective is not viable for web-based scholarship, where nodes must interoperate in order to arrive at a thriving web-based scholarly ecology. In this sense, the desired future of web-based scholarly communication and research more closely resembles the mash-up vision of Web 2.0 than the monopolistic and centralistic tendency of many current web portals. We think that the REST/HATEOAS principles, as applied in the thinking of Signposting the Scholarly Web, can provide the unifying framework to knit scholarly nodes together.

Over the years, we have learned that no one is "King of Scholarly Communication" and that no progress regarding interoperability can be accomplished without active involvement and buy-in from the stakeholder communities. However, it is a significant challenge to determine what exactly the stakeholder communities are, and who can act as their representatives, when the target environment is as broad as all nodes involved in web-based scholarship. To put this differently, it is hard to know how to exactly start an effort to work towards increased interoperability. Let us know, if you have ideas regarding a way forward.

Acknowledgements

The projects discussed involved numerous collaborators, acknowledged in the respective specifications. It has been a pleasure to work with them and we hope that our recollections of the last 15 years are consistent with theirs.

References

[1] Bechhofer, S., Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, and Carole Goble. 2013. "Why linked data is not enough for scientists." *Future Generation Computer Systems* 29 (2) 599-611. http://dx.doi.org/10.1016/j.future.2011.08.004

[2] Bekaert, J. and Herbert Van de Sompel. 2005. "A standards-based solution for the accurate transfer of digital assets." *D-Lib Magazine* 11 (6). http://dx.doi.org/10.1045/june2005-bekaert

[3] Carroll, J. J., Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. "idd graphs, provenance and trust." *Proceedings of the 14th International Conference on World Wide Web* pp 613-622. http://dx.doi.org/10.1145/1060745.1060835

[4] Fielding, R. T. 2000. "Architectural styles and the design of network-based software architectures." *Doctoral dissertation, University of California, Irvine* https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[5] Granell, G., Laura Díaz, Sven Schade, Nicole Ostländer, and Joaquín Huerta. 2013. "Enhancing integrated environmental modelling by designing resource-oriented interfaces." *Environmental Modelling & Software* 39(January) 229-246. http://dx.doi.org/10.1016/j.envsoft.2012.04.013

[6] Hagedorn, K. and Joshua Santelli. 2008. "Google Still Not Indexing Hidden Web URLs." *D-Lib Magazine* 14 (7/8). http://dx.doi.org/10.1045/july2008-hagedorn

[7] Masanes, J. 2006. "Web archiving." *Springer Verlag* ISBN 978-3-540-23338-1 http://dx.doi.org/10.1007/978-3-540-46332-0

[8] McCown, F., Xiaoming Liu, Michael L. Nelson, and Mohammad Zubair. 2006. "Search Engine Coverage of the OAI-PMH Corpus." *IEEE Internet Computing* 10 (2) 66-73. http://dx.doi.org/10.1109/MIC.2006.41

[9] Page, K. R., David De Roure, and Kirk Martinez. 2011. "REST and Linked Data: a match made for domain driven development?" *Proceedings of the Second International Workshop on RESTful Design at WWW 2011* http://www.ws-rest.org/2011/proc/a5-page.pdf

[10] Roosendaal, H., and Peter Geurts. 1997. "Forces and functions in scientific communication: an analysis of their interplay." *Cooperative Research Information Systems in Physics, August 31—September 4 1997, Oldenburg, Germany* http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html

[11] Smith, J., and Michael L. Nelson. 2008. "Site Design Impact on Robots: An Examination of Search Engine Crawler Behavior at Deep and Wide Websites." *D-Lib Magazine* 14 (3/4). http://dx.doi.org/10.1045/march2008-smith

[12] Van de Sompel, H., Thomas Krichel, Patrick Hochstenbach, Victor M. Lyapunov, Kurt Maly, Mohammad Zubair, Mohammad Kholief, and Michael L. Nelson. 2000. "The UPS prototype." *D-Lib Magazine* 6 (2). http://dx.doi.org/10.1045/february2000-vandesompel-ups

[13] Van de Sompel, H., Sandy Payette, John Erickson, Carl Lagoze, and Simeon Warner. 2004. "Rethinking scholarly communication: Building the system that scholars deserve." *D-Lib Magazine* 10 (9). http://dx.doi.org/10.1045/september2004-vandesompel

[14] Van de Sompel, H., Michael L. Nelson, Carl Lagoze, and Simeon Warner. 2004. "Resource Harvesting within the OAI-PMH Framework." *D-Lib Magazine* 10 (12). http://dx.doi.org/10.1045/december2004-vandesompel

[15] Van de Sompel, H., Michael L. Nelson, Robert Sanderson, Lyudmila L. Balakireva, Scott Ainsworth, and Harihar Shankar. 2009. "Memento: Time Travel for the Web." *arXiv.org* http://arxiv.org/abs/0911.1112

[16] Van de Sompel, H., Robert Sanderson, Harihar Shankar, and Martin Klein. 2014. "Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping." *The International Journal of Digital Curation* 9(1) 331-342. http://dx.doi.org/10.2218/ijdc.v9i1.320

About the Authors

Herbert Van de Sompel is an information scientist at the Los Alamos National Laboratory and leads the Prototyping Team in the Research Library. The Team does research regarding various aspects of scholarly communication in the digital age. Herbert has played a role in various interoperability efforts (OAI-PMH, OpenURL, OAI-ORE, info URI, Open Annotation, ResourceSync, SharedCanvas, Memento) and in the design of scholarly discovery tools (SFX linking server, bX recommender engine).Currently, he works on the Robust Links effort, aimed at addressing reference rot, and contemplates about Archiving the Web-Based Scholarly Record. More information about Herbert can be found at http://public.lanl.gov/herbertv/.

Michael L. Nelson is a professor of computer science at Old Dominion University. Prior to joining ODU, he worked at NASA Langley Research Center from 1991-2002. He is a co-editor of the OAI-PMH, OAI-ORE, Memento, and ResourceSync specifications. His research interests include repository-object interaction and alternative approaches to digital preservation. More information about Dr. Nelson can be found at http://www.cs.odu.edu/~mln/.