Signposting for Repositories

Martin Klein
Research Library
Los Alamos National Laboratory
Los Alamos, NM, USA
<http://orcid.org/0000-0003-0130-2097>
mklein@lanl.gov

Harihar Shankar
Research Library
Los Alamos National Laboratory
Los Alamos, NM, USA
<http://orcid.org/0000-0003-4949-0728>
harihar@lanl.gov

Herbert Van de Sompel
Research Library
Los Alamos National Laboratory
Los Alamos, NM, USA
<http://orcid.org/0000-0002-0715-6126>
herbertv@lanl.gov

ABSTRACT

Digital repositories can often easily be navigated by humans but not by machines. We introduce Signposting, a mechanism to show machines how to maneuver repositories' objects and how to interpret their relationships. Signposting is based on standard and widely adopted web technologies - typed links and HTTP link headers.

Keywords

Typed Links, Machine-Friendly Repositories, REST, HATEOS

1. INTRODUCTION

Modern digital repositories offer an impressive set of functionalities. Storage options for a variety of file types, APIs to allow for access to stored objects, fixity checking, versioning, and audit are just a few of the services frequently provided. However, while most of the common repository systems are accessible via the web, they are often lacking in adoption of common web standards. In other words, modern repositories are mostly **on the web** but not truly **of the web**. This situation becomes apparent when illustrated by an example in the realm of scholarly repositories. For a human, it is arguably fairly simple to navigate through a repository's holdings and discover an object, its core content, auxiliary content, bibliographic description, persistent identifier, and author names/identifiers. Often all of this data can be obtained without much effort by simply scanning the object's landing page. For a machine, however, the same task is much harder to accomplish, in part due to repository-specific conventions as to how to convey this information. This situation requires developers to implement heuristics in order to

"interpret" the data surfaced by repositories - a proposition that does not scale and hence hinders the emergence of 3rd party value-added services in the scholarly communication landscape.

In this poster we outline the Signposting concept as a simple yet potentially powerful solution to this problem. Signposting is a standards-based approach for digital repositories in libraries, archives, and museums to expose data in a machine-friendly manner and to clarify common patterns and relationships between resources they hold.

2. SIGNPOSTING PATTERNS

The core intention of Signposting¹ is to make it easier for applications to **navigate** across repository resources, not to actually **describe** them. To that end, Signposting uses a REST/HATEOS approach that consists of conveying typed links² in HTTP Link response headers³ and/or the <link> element in HTML pages. These standardized technologies are easy to implement and therefore frequently used on the web. For example, motivated by LOCKSS⁴ and CLOCKSS⁵ use cases, Signposting has previously been explored to support preservation efforts for e-Journals [Van de Sompel et al. 2016b]. Here we focus on the following, more generic but very typical usage patterns of scholarly repositories:

The *Object Identifier* pattern serves the purpose of helping machines find the identifying HTTP URI of an object. Figure 1a depicts a DOI redirecting to a landing page of a scholarly object. The landing page also links to resources that are part of the publication, such as alternate renderings of the article, all of which are identified by the same DOI. A machine that only discovers this landing page, for example by following a citation that used the URI of the landing page instead of the DOI, a scenario frequently observed [Van de Sompel et al. 2016a], has no easy way to determine the identifying URI (the DOI) of the scholarly object. Similarly, it also has no convenient way of obtaining the DOI if it only

¹<http://signposting.org/>

²<https://www.iana.org/assignments/link-relations/link-relations.xhtml>

³<https://tools.ietf.org/html/rfc5988>

⁴<https://www.lockss.org/>

⁵<https://clockss.org/>

discovered the PDF version of the article, for example. If, however, the repository implemented an HTTP link header pointing from the landing page and the publication representations at the DOI URI using the relation type “cite-as”⁶, a machine would be able to figure out that the article and its related resources are identified by the same DOI.

The *Author Information* pattern, depicted in Figure 1b, enables machines to discover author information of repository objects. The scenario is similar to the one described above: a link is implemented with the “author” relation type⁷ from the landing page and the HTTP DOI to the author identifying URI, for example, the author’s ORCID URI. The below HTTP link headers that are returned in response to a HEAD/GET on the landing page convey object and author identifiers for the resource <https://doi.pangaea.de/10.1594/PANGAEA.880814> hosted by the Pangaea service, one of the early adopters of the Signposting concept.

```
HTTP/1.1 200 OK
Server: PANGAEA/1.0
Date: Wed, 31 Jan 2018 05:58:42 GMT
Link: <https://doi.org/10.1594/PANGAEA.880814>;
rel="cite-as",
<https://orcid.org/0000-0001-5212-9808>;rel="author",
<https://orcid.org/0000-0002-9017-0083>;rel="author"
```

Bibliographic Metadata can be made discoverable by adopting the pattern shown in Figure 1c. A DOI redirects to the landing page which in turn links to bibliographic resources. If these two resources provide link headers with the relation type “describedby”⁸ pointing at bibliographic resources provided by CrossRef or DataCite, a machine can easily make that connection between the DOI-identified resource and its metadata records. To make these links even more expressive and friendly to machines, they should also have a “type” attribute to convey the format of the linked metadata record. In reverse, links to the HTML page and to the HTTP DOI URI with the relation type “describes”⁹ from the corresponding bibliographic resources can also be provided to support bidirectional navigation.

The *Boundary Information* pattern is related to automatically expressing the web boundary of an object. Frequently, landing pages link to related resources such as alternate renderings of the article (as seen above) and supplemental information. Clearly, these resources should be considered as being within the boundary of a publication. However, landing pages also link to other external resources such as advertisement or social media platforms, which, in turn, are not within the publication’s boundary. The distinction between what falls within and what outside of the boundary is trivial for a human to make. For a machine, however, this is very hard to delineate. This pattern entails the landing page providing link headers with the relation type “item”¹⁰ to point to in-boundary resources. In turn, these resources provide links with the relation type “collection”¹¹ to point back to the landing page.

The Signposting concept is not limited to these four patterns. Other examples include the expression of license in-

⁶<https://www.ietf.org/id/draft-vandesompel-citeas-02.txt>

⁷<https://www.w3.org/TR/html5/links.html>

⁸<https://tools.ietf.org/html/rfc6892>

⁹<https://tools.ietf.org/html/rfc6892>

¹⁰<https://tools.ietf.org/html/rfc6573>

¹¹<https://tools.ietf.org/html/rfc6573>

Figure 1: Signposting patterns

formation with the “license” link relation type¹² or exposing object versions with relation types such as “previous memento” and “next memento”¹³.

3. ACKNOWLEDGMENTS

This work is in part supported by the Andrew W. Mellon Foundation, grant number 11600663.

4. CONCLUSION

¹²<https://tools.ietf.org/html/rfc4946>

¹³<https://tools.ietf.org/html/rfc7089>

Signposting is a standard-based approach that follows the REST/HATEOS “follow-your-nose” principle to make it easier for machines to navigate across repository resources. We outline the Signposting concept on four common usage patterns and detail which link relation types should be used for the corresponding patterns. Adopting the Signposting approach helps repositories improve their web integration and increase their level of discoverability and interoperability with other systems. Ultimately, making repositories more friendly to machines will lead to applications that are beneficial for humans.

5. REFERENCES

- [Van de Sompel et al. 2016a] Herbert Van de Sompel, Martin Klein, and Shawn M. Jones. 2016a. Persistent URIs Must Be Used To Be Persistent. In *Proceedings of WWW '16*. 119–120.
<https://doi.org/10.1145/2872518.2889352>
- [Van de Sompel et al. 2016b] Herbert Van de Sompel, David Rosenthal, and Michael L. Nelson. 2016b. Web Infrastructure to Support e-Journal Preservation (and More). <http://arxiv.org/abs/1605.06154>.