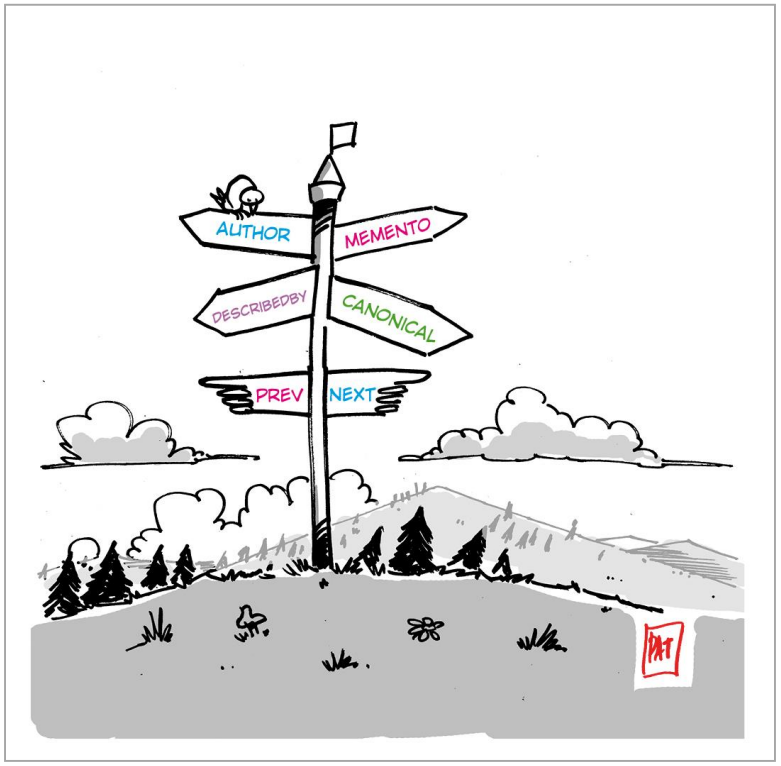
Signposting the Scholarly Web: An Overview

http://signposting.org



Cartoon by Patrick Hochstenbach

Slides prepared by:
Herbert Van de Sompel
@hvdsomp

Acknowledgments: Geoff Bilder, Shawn Jones, Martin Klein, Michael L. Nelson, David Rosenthal, Harihar Shankar, Simeon Warner, Karl Ward, Joe Wass

Signposting is funded by the Andrew W. Mellon Foundation

Outline

- **Introduction: HTTP Links**
- Signposting the Scholarly Web
- Proposed Patterns
- Putting it Together

HTTP Links

Internet Engineering Task Force (IETF)
Request for Comments: 5988
Updates: 4287
Category: Standards Track
ISSN: 2070-1721

M. Nottingham
October 2010

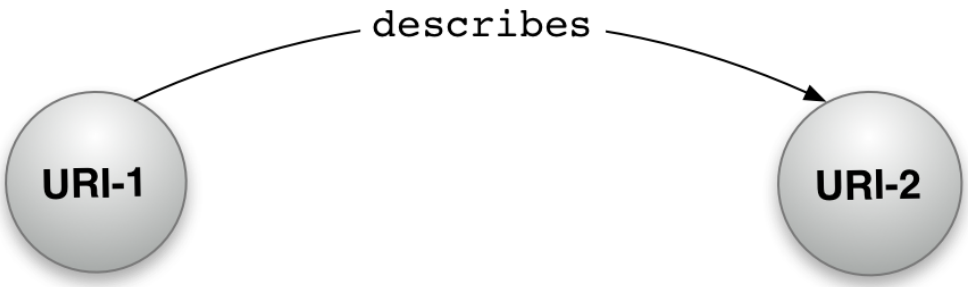
Web Linking

Abstract

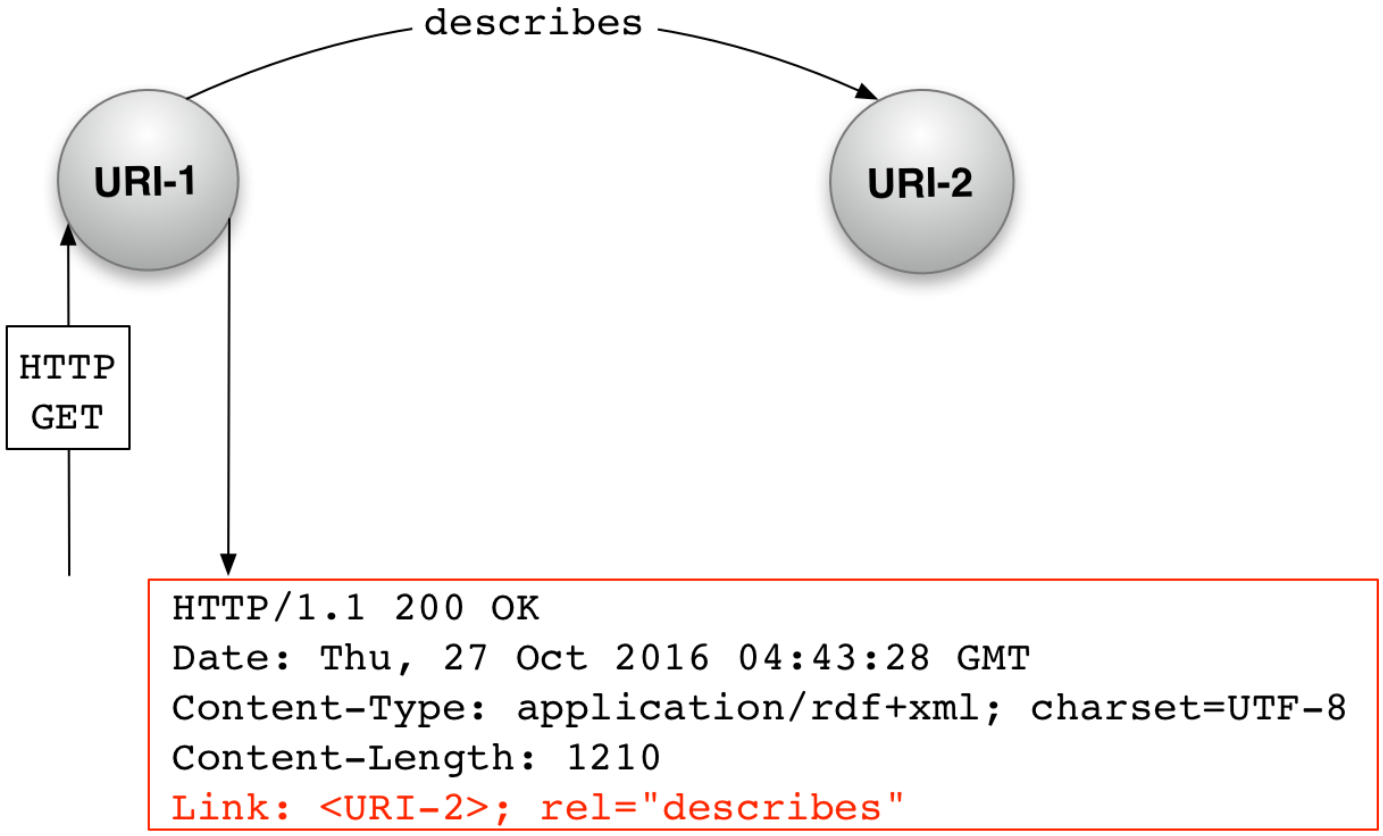
This document specifies relation types for Web links, and defines a registry for them. It also defines the use of such links in HTTP headers with the Link header field.

Mark Nottingham (2010) RFC5988: Web Linking.
http://tools.iets.org/rfc/rfc5988.txt

HTTP Links



HTTP Links



HTTP Links Are Used

```
curl –I http://dbpedia.org/data/Reykjavik

HTTP/1.1 200 OK
Date: Thu, 27 Oct 2016 04:43:28 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 1210
Link:
 <http://creativecommons.org/licenses/by-sa/3.0>
 ; rel=“license",
 <http://dbpedia.org/data/Reykjavik>
 ; rel="alternate"; type="text/n3",
 <http://dbpedia.org/resource/Reykjavik>; rel="describes",
 <http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/
  data/Reykjavik>
 ; rel="timegate"
```

HTTP Links Are Used

```
curl -I http://dbpedia.org/data/Reykjavik

HTTP/1.1 200 OK
Date: Thu, 27 Oct 2016 04:43:28 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 1210
Link:
 <http://creativecommons.org/licenses/by-sa/3.0>
 ; rel="license",
 <http://dbpedia.org/data/Reykjavik>
 ; rel="alternate"; type="text/n3",
 <http://dbpedia.org/resource/Reykjavik>; rel="describes",
 <http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/
  data/Reykjavik>
 ; rel="timegate"
```

HTTP Links Are Used

```
curl -I http://dbpedia.org/data/Reykjavik

HTTP/1.1 200 OK
Date: Thu, 27 Oct 2016 04:43:28 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 1210
Link:
 <http://creativecommons.org/licenses/by-sa/3.0>
 ; rel="license",
 <http://dbpedia.org/data/Reykjavik>
 ; rel="alternate"; type="text/n3",
 <http://dbpedia.org/resource/Reykjavik>; rel="describes",
 <http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/
  data/Reykjavik>
 ; rel="timegate"
```

HTTP Links Are Used

```
curl -I http://dbpedia.org/data/Reykjavik

HTTP/1.1 200 OK
Date: Thu, 27 Oct 2016 04:43:28 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 1210
Link:
 <http://creativecommons.org/licenses/by-sa/3.0>
 ; rel="license",
 <http://dbpedia.org/data/Reykjavik>
 ; rel="alternate"; type="text/n3",
 <http://dbpedia.org/resource/Reykjavik>; rel="describes",
 <http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/
  data/Reykjavik>
 ; rel="timegate"
```

HTTP Links Are Used

```
curl -I http://dbpedia.org/data/Reykjavik

HTTP/1.1 200 OK
Date: Thu, 27 Oct 2016 04:43:28 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 1210
Link:
 <http://creativecommons.org/licenses/by-sa/3.0>
 ; rel="license",
 <http://dbpedia.org/data/Reykjavik>
 ; rel="alternate"; type="text/n3",
 <http://dbpedia.org/resource/Reykjavik>; rel="describes",
 <http://mementoarchive.lanl.gov/dbpedia/timegate/http://dbpedia.org/
  data/Reykjavik>
 ; rel="timegate"
```

HTTP Link Relation Types

- Registered in IANA registry
  - Strings, e.g. `license`, `alternate`, `describes`, `timegate`
  - Requires a formal specification, e.g. RFC
  - Typically used for common relationships, generically specified
  - Provides broad, coarse grained interoperability

- Minted by a community
  - URIs, e.g. `http://xmlns.com/foaf/0.1/primaryTopic`
  - Requires community agreement
  - Can be as specific as desired
  - Can provide community-specific, fine grained interoperability

HTTP Links Are Pretty Neat

- Can uniformly be used for all MIME types
- Accessible via HTTP HEAD (no content transfer):
  - Works for large resources and for restricted content
- HTTP Links can be conveyed:
  - by-value, in the HTTP Link header
  - by-reference, by using a `linkset` link in the HTTP header that points to a collection of links (1)
- HTTP Links provide guidance to machine agents intent on accomplishing a specific task

---

(1) Wilde, E. and Van de Sompel, H (2017) Linkset: A Link Relation Type and Media Types for Link Sets https://datatracker.ietf.org/doc/draft-wilde-linkset-link-rel/

Outline

- Introduction: HTTP Links
- **Signposting the Scholarly Web**
- Proposed Patterns
- Putting it Together

Signposting the Scholarly Web

- Proposal:

  Use HTTP Links to address some long standing problems regarding scholarly resources on the web, by interlinking them using appropriate relation types

- Focus on a limited set of patterns to support uniformly:
  - **Conveying a Persistent Identifier**
  - **Expressing the web boundary of a scholarly resource**
  - **Making bibliographic metadata discoverable**
  - Conveying an Author Identifier
  - Conveying a resource type

Terminology

- **PID**: Persistent Identifier
- **HTTP PID**: the HTTP URI *notation* of the PID
- **entry page**: the page where one ends up after following redirects from the HTTP PID, typically the landing page or full content HTML
- **resource**: a web resource identified by an HTTP URI
- **constituent resource**: a resource that is an integral part of a scholarly object, e.g. landing page, PDF, supporting data, …
- **bibliographic resource**: a resource that provides a bibliographic description of a scholarly object

Outline

- Introduction: HTTP Links
- Signposting the Scholarly Web
- **Proposed Patterns**
- Putting it Together

Pattern: Identifier

- Problem: When positioned at a constituent resource of a scholarly object, the associated HTTP PID is not available. As a result:
  - Landing page URIs are used for citation (1)
  - Applications such as annotation can not determine the HTTP PID associated with a constituent resource

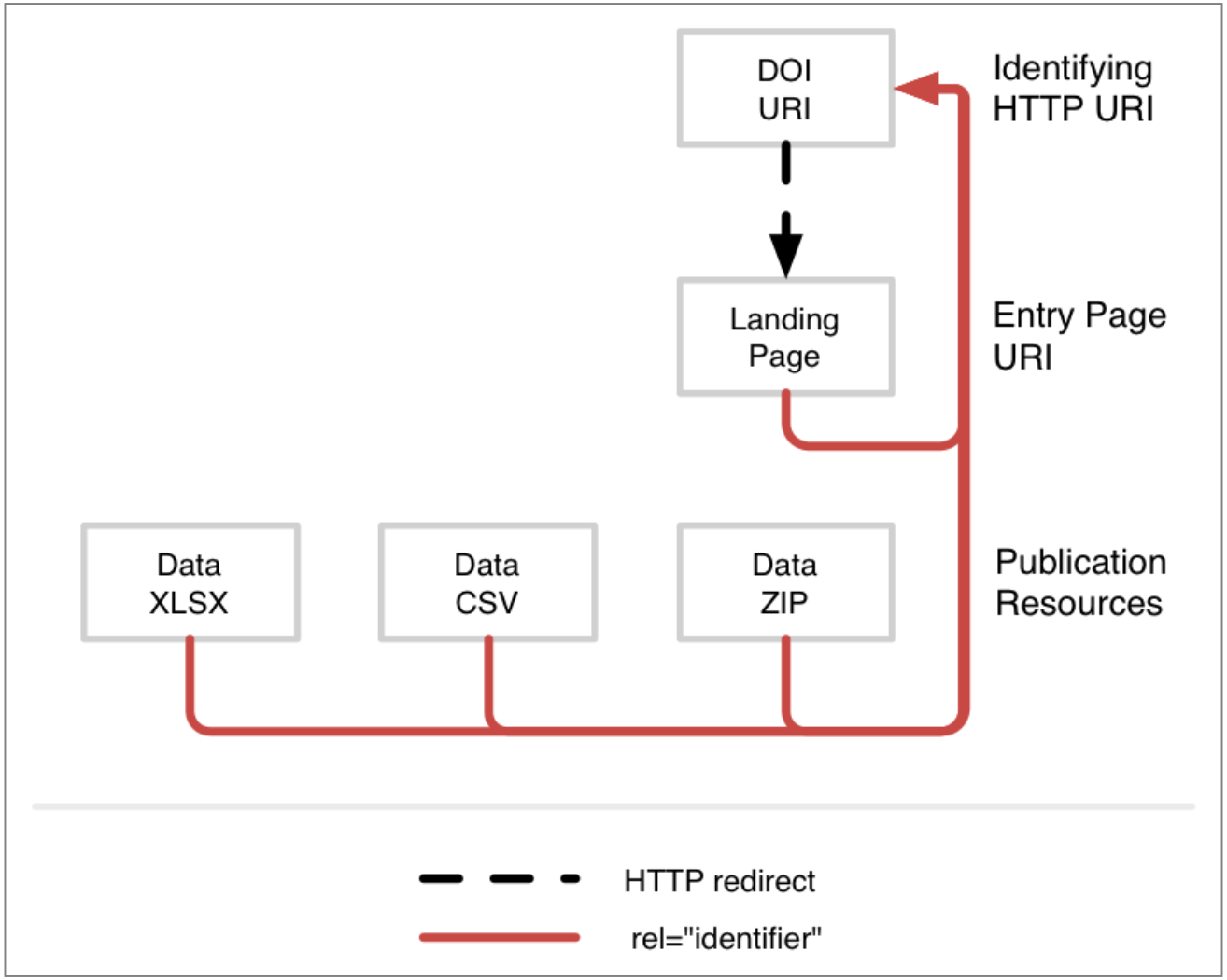
- Solution: provide `identifier` link pointing at the HTTP PID

- Applies to: **entry page**, all constituent resources

---

(1) Herbert Van de Sompel, Martin Klein, and Shawn Jones (2016) Persistent URIs Must Be Used to Be Persistent. In: WWW2016. http://arxiv.org/1602.09102

Use HTTP Link with `identifier` Relation Type



http://signposting.org/identifier/dryad/

Use HTTP Link with `identifier` Relation Type

```
curl -I
http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html


HTTP/1.1 200 OK
Date: Wed, 26 Oct 2016 12:36:37 GMT
Server: Apache/2.2.15 (CentOS)
Last-Modified: Thu, 19 Nov 2015 14:50:19 GMT
ETag: "205a5e-f5ef-524e5e0ab80c0"
Accept-Ranges: bytes
Content-Length: 62959
Content-Type: text/html; charset=UTF-8
Link: <https://doi.org/10.1045/november2015-vandesompel>
 ; rel="identifier"
```

Pattern: Publication Boundary

- Problem: It is not possible to determine what the constituent resources of a scholarly object are
  - Preservation and text mining tools require portal-specific heuristic to find those constituent resources (1)
  - Can’t find the pathway from an HTTP PID directly to e.g. the PDF

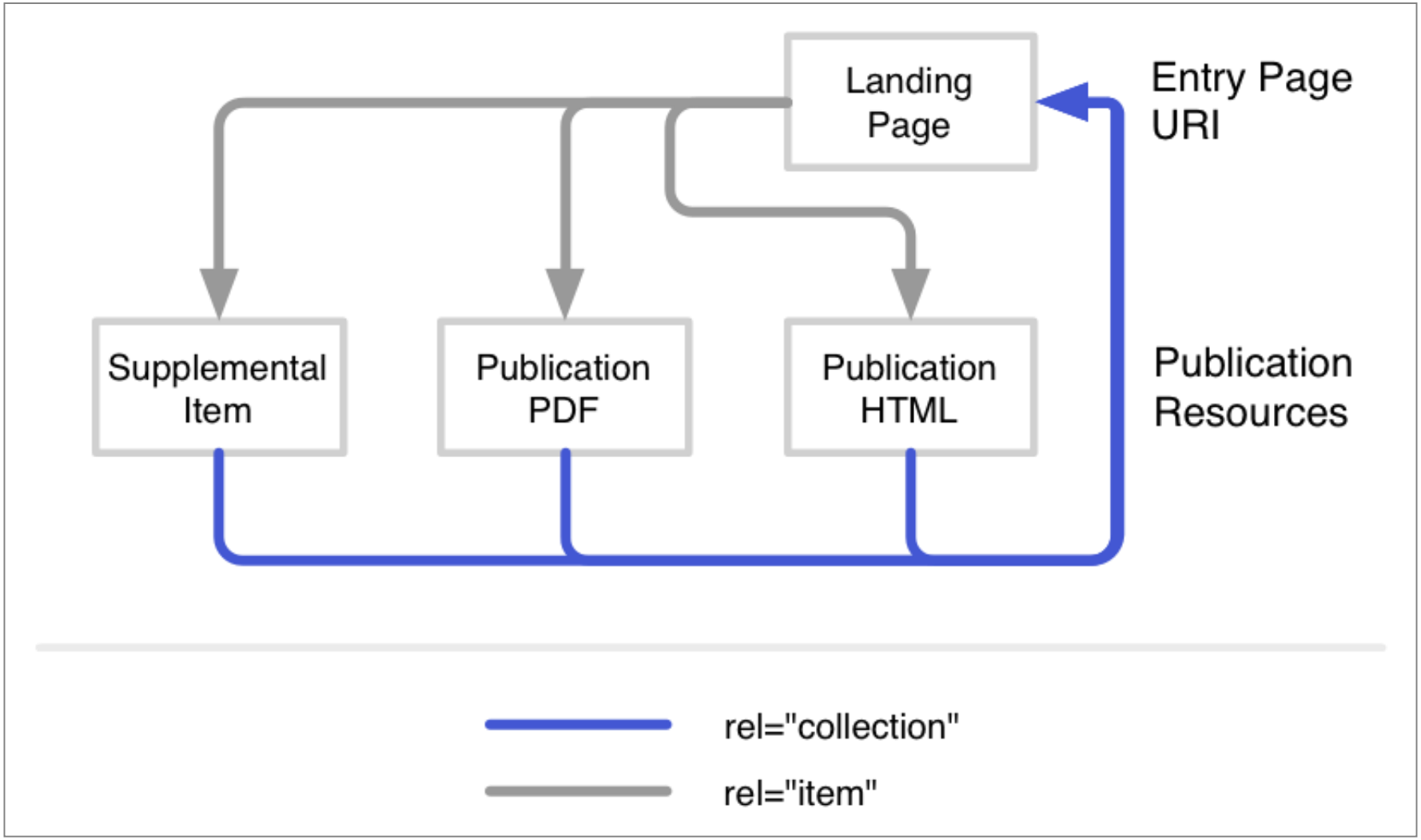
- Solution: provide `item/collection` links to interlink entry page and constituent resources; convey MIME types on `item` links

- Applies to: **All constituent resources of a scholarly object**

---

(1) Van de Sompel, H., Rosenthal, D., and Nelson, M.L. (2016) Web Infrastructure to Support e-Journal Preservation (and More). http://arxiv.org/abs/1605.06154

Use HTTP Link with `item/collection` Relation Type



http://signposting.org/publication_boundary/oxford/

Use HTTP Link with `item/collection` Relation Type

```
curl -I http://www.irrodl.org/index.php/irrodl/article/view/2179


HTTP/1.1 200 OK
Date: Sat, 17 Jun 2017 09:11:52 GMT
Server: Apache
Vary: Accept-Encoding
Accept-Ranges: none
Keep-Alive: timeout=5, max=200
Content-Type: text/html; charset=utf-8
Link:
 <http://www.irrodl.org/index.php/irrodl/article/download/2179/3748>
 ; rel="item" ; type="application/pdf" ,
<http://www.irrodl.org/index.php/irrodl/article/view/2179/3747>
 ; rel="item"; type="text/epub+zip"
```

Pattern: Bibliographic Metadata

- Problem: It is not possible to determine where the bibliographic resources that describes a scholarly object can be found
  - Preservation and reference manager tools require portal-specific heuristic to find those resources (1)

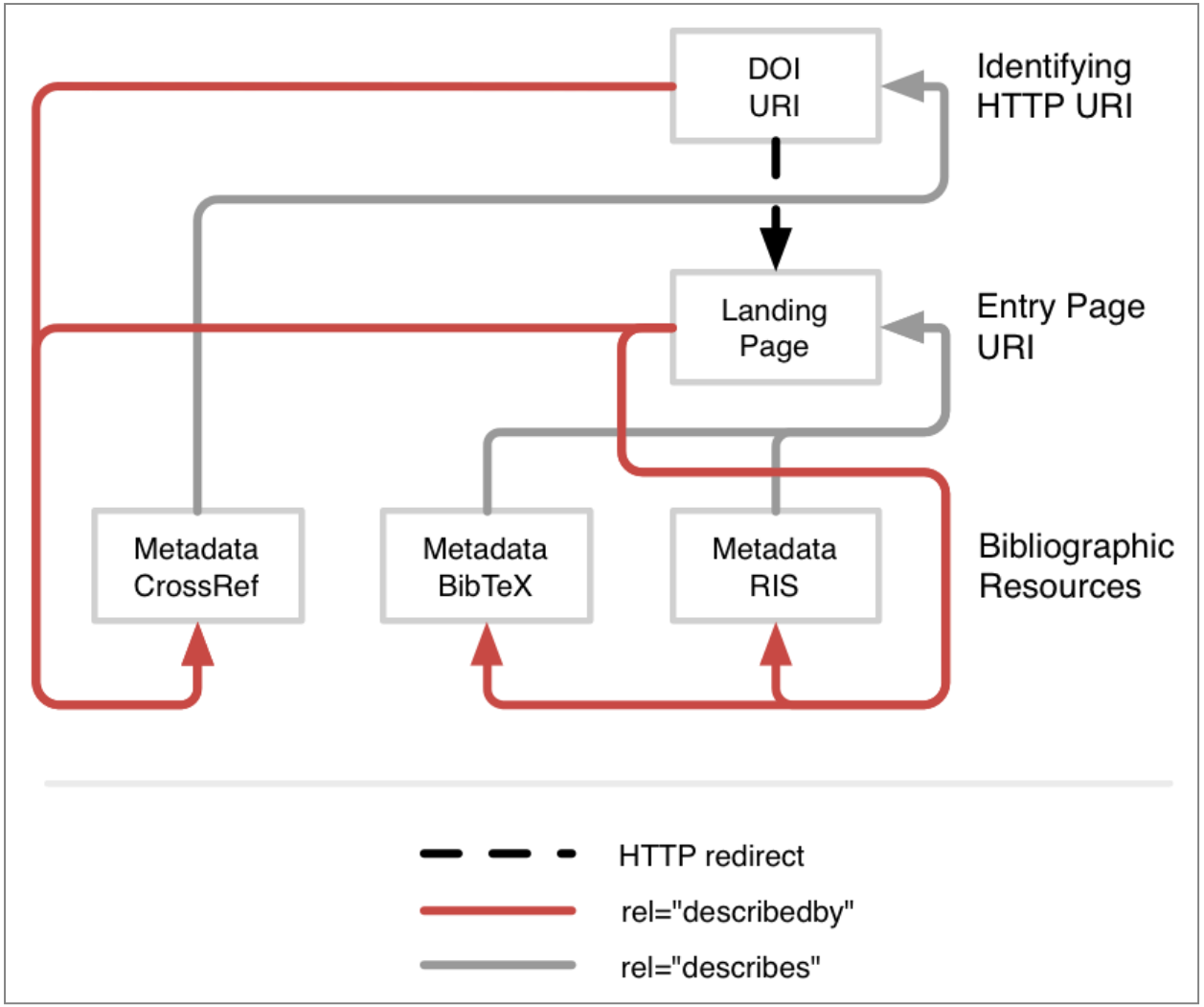
- Solution: provide `describedby/describes` links to interlink entry page and bibliographic metadata resources

- Applies to:
  - `describedby`: **HTTP PID**, **entry page**
  - `describes`: bibliographic resources

---

(1) Van de Sompel, H., Rosenthal, D., and Nelson, M.L. (2016) Web Infrastructure to Support e-Journal Preservation (and More). http://arxiv.org/abs/1605.06154

Use HTTP Link with `describedby/describes` Relation Type



http://signposting.org/bibliographic_metadata/springer/

Use HTTP Link with `describedby/describes` Relation Type

```
curl -I
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.01152
53

HTTP/1.1 200 OK
Date: Sat, 17 Jun 2017 09:29:03 GMT
Server: Apache-Coyote/1.1
Content-Language: en-US
Keep-Alive: timeout=5, max=100
Content-Length: 308491
Content-Type: text/html;charset=UTF-8
Connection: keep-alive
Link:
 <http://journals.plos.org/plosone/article/citation/bibtex?id=10.1371
 %2Fjournal.pone.0115253>
 ; rel="describedby" ; type="application/x-bibtex" ,
 <https://doi.org/10.1371/journal.pone.0115253>
 ; rel="describedby" ; type="application/vnd.citationstyles.csl+json"
```

Bibliographic Metadata Conventions

Format	MIME Type
BibTeX	`application/x-bibtex`
CiteProc JSON	`application/vnd.citationstyles.csl+json`
RIS	`application/x-research-info-systems`

Many other bibliographic formats are in use and many share `text/plain`, `application/xml`, or `application/json` as MIME types. In order to distinguish between formats of a same MIME type, either a dedicated MIME type should be registered or a `profile` attribute can be used on a `describedby` link to clarify the format beyond its MIME type. The former allows for content negotiation for the format, the latter does not. The below table shows how the `profile` approach can be used for popular XML-based formats.

Format	MIME Type	`profile` Attribute Value
MARC XML	`application/xml`	`http://www.loc.gov/MARC21/slim`
MODS	`application/xml`	`http://www.loc.gov/mods/`
Simple Dublin Core	`application/xml`	`http://purl.org/dc/elements/1.1/`
Qualified Dublin Core	`application/xml`	`http://dublincore.org/documents/dcmi-terms/`

http://signposting.org/conventions/

Outline

- Introduction: HTTP Links
- Signposting the Scholarly Web
- Proposed Patterns
- **Putting it Together**


HTTP
DOI


{JSON}


Landing
Page


TEX
Bib
TEX


CSV
1,2,3

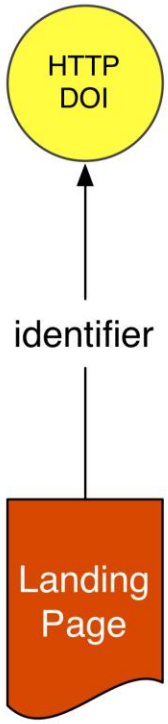

HTML


PDF
Adobe

HTTP
PID

bibliographic
resources

constituent
resources



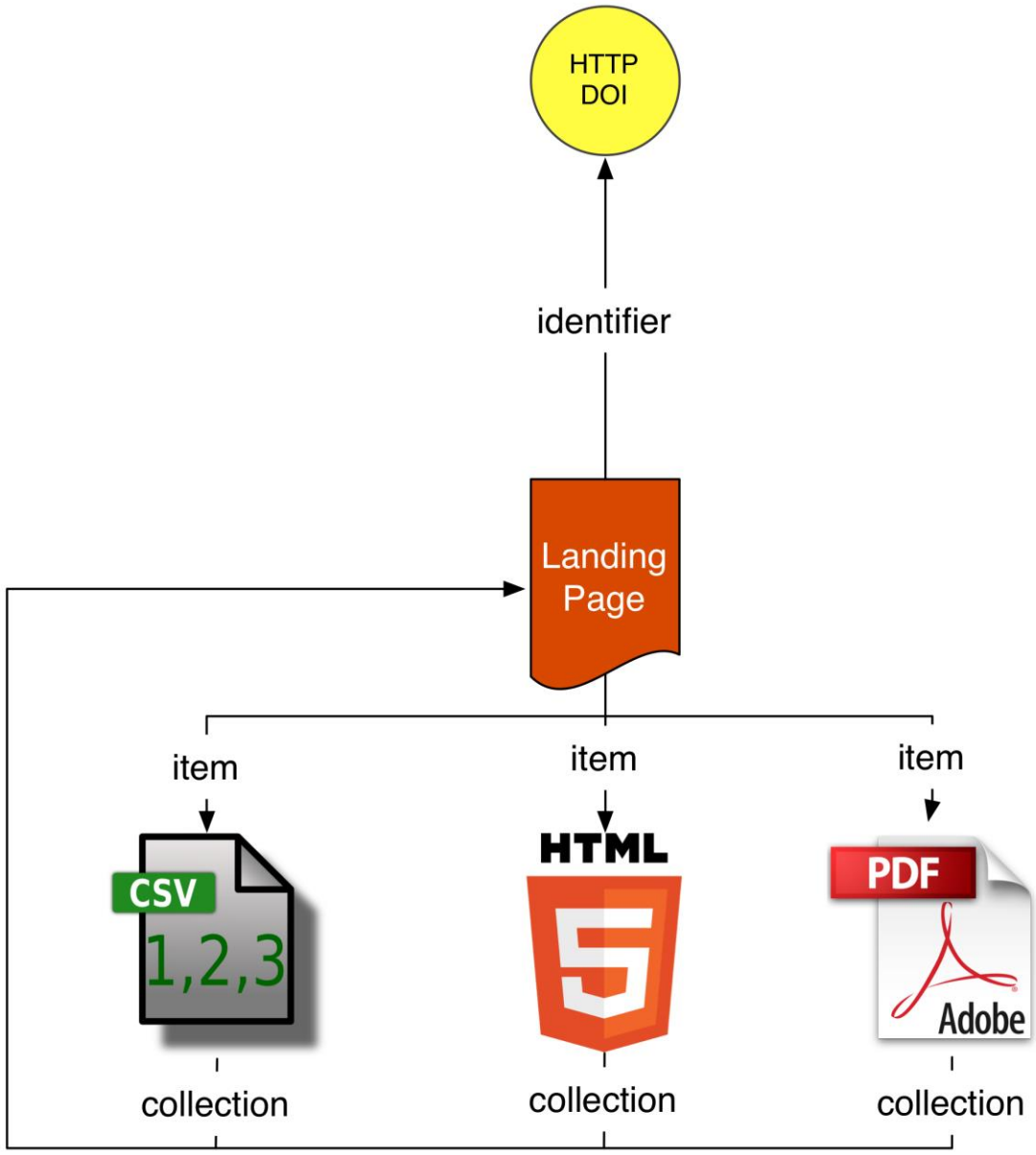
{JSON}

BibTeX

CSV
1,2,3

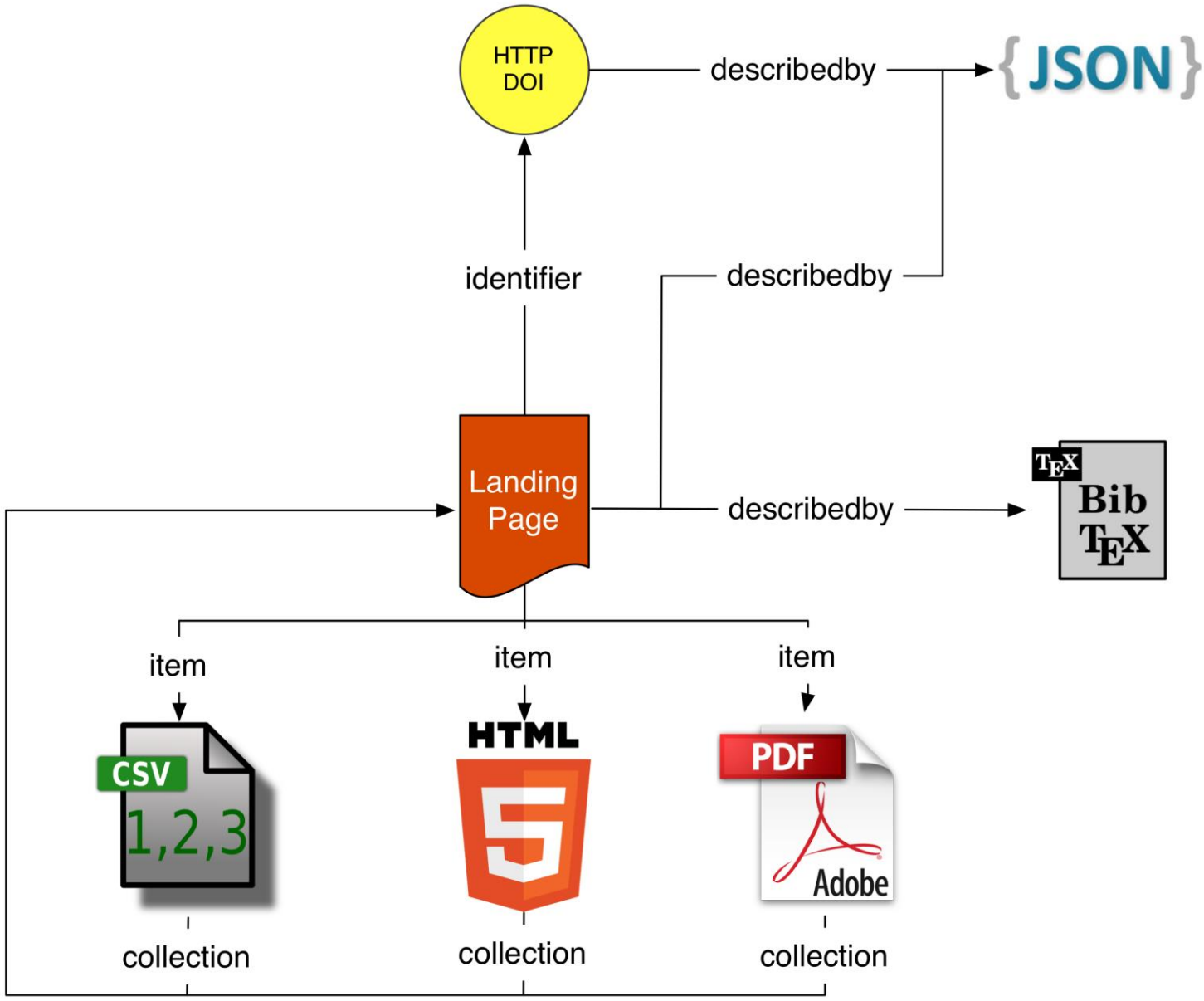
HTML

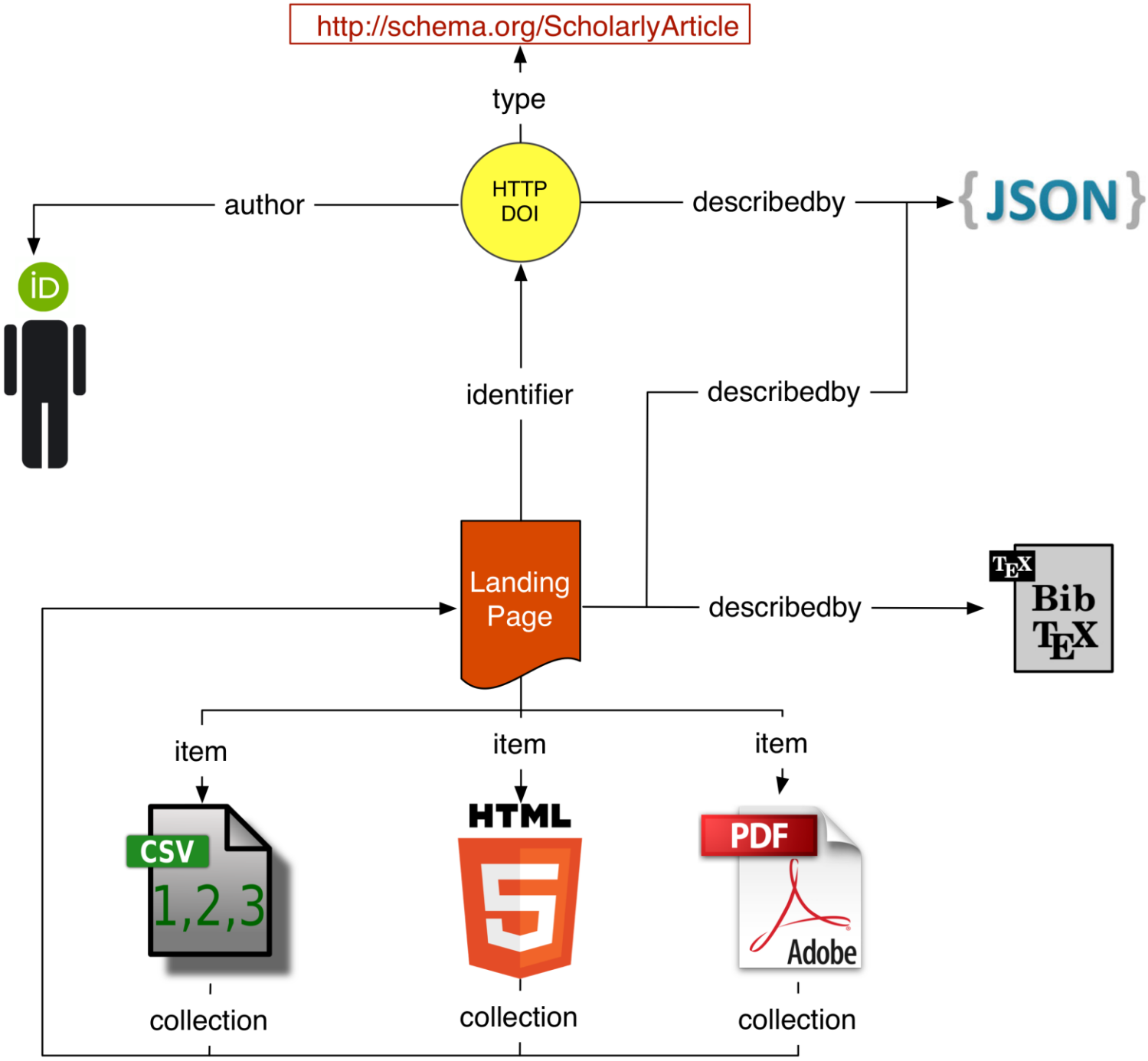
PDF
Adobe


HTTP
DOI
identifier
Landing
Page
item
item
item
CSV
1,2,3
HTML
PDF
Adobe
collection
collection
collection

{JSON}

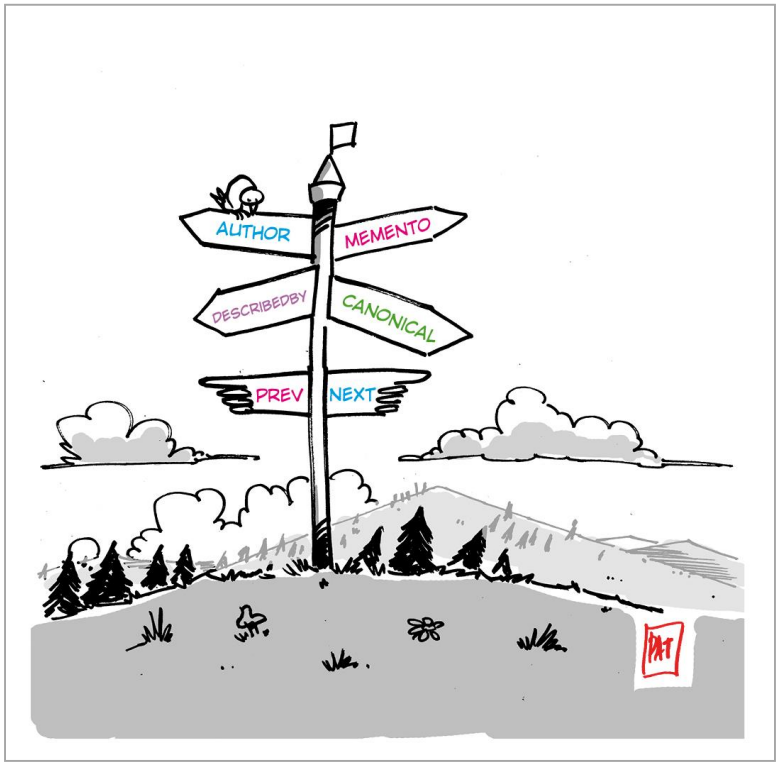
BibTeX


HTTP
DOI
describedby
{JSON}
identifier
describedby
Landing
Page
describedby
TEX
Bib
TEX
item
item
item
CSV
1,2,3
HTML
PDF
Adobe
collection
collection
collection


http://schema.org/ScholarlyArticle
type
HTTP
DOI
author
describedby
{JSON}
identifier
describedby
Landing
Page
describedby
Bib
TEX
item
item
item
CSV
1,2,3
HTML
PDF
Adobe
collection
collection
collection

Signposting the Scholarly Web: An Overview

http://signposting.org



Cartoon by Patrick Hochstenbach

Slides prepared by:
Herbert Van de Sompel
@hvdsomp

Acknowledgments: Geoff Bilder, Shawn Jones, Martin Klein, Michael L. Nelson, David Rosenthal, Harihar Shankar, Simeon Warner, Karl Ward, Joe Wass

Signposting is funded by the Andrew W. Mellon Foundation