Specifications of Next Generation Repositories

Towards a Scholarly Commons

MS15 – Specifications of Next Generation Repositories
Version 1.0 – Final
PUBLIC

H2020-EINFRA-2017
Grant Agreement 777541
Document Description

MS15 – Specifications of Next Generation Repositories

WP6 – Towards a Scholarly Commons

WP participating organizations: UNIBI, UGOE, CNR, ATHENA RC, UMINHO, EIFL, CERES, COAR E.V., CLARA, KNAW, UEDIN, NHRF (EIE), OLH, RBI, JISC

Contractual Delivery Date: 12/2018
Actual Delivery Date: 01/2019

Nature: Report
Version: 1.0

Public Deliverable

Preparation Slip

<table>
<thead>
<tr>
<th>Name</th>
<th>Organisation</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>José Carvalho, Eloy Rodrigues, Kathleen Shearer</td>
<td>UMinho, COAR</td>
</tr>
<tr>
<td>Edited by</td>
<td>Pedro Príncipe</td>
<td>UMINHO</td>
</tr>
<tr>
<td>Reviewed by</td>
<td>Pedro Príncipe</td>
<td>UMINHO</td>
</tr>
</tbody>
</table>

Revision History

<table>
<thead>
<tr>
<th>Issue</th>
<th>Item</th>
<th>Reason for Change</th>
<th>Author</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0.1</td>
<td>First Delivery</td>
<td>November 2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1.0</td>
<td>First version</td>
<td>January 2019</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table of Contents

1. INTRODUCTION 5

2. RESOURCESYNC 6
   2.1 USE CASES 6
   2.2 IMPLEMENTING RESOURCESYNC IN YOUR REPOSITORY 8
   2.3 SERVICE PROVIDERS USING RESOURCESYNC 8

3. SIGNPOSTING 9
   3.1 USE CASES 9
   3.1.1 USE CASES BY SERVICE 10
   3.2 SIGNPOSTING PATTERNS 10
      3.2.1 AUTHOR – HTTPS://SIGNPOSTING.ORG/AUTHOR/ 10
      3.2.2 BIBLIOGRAPHIC METADATA 11
      3.2.3 IDENTIFIER 11
      3.2.4 PUBLICATION BOUNDARY 11
      3.2.5 RESOURCE TYPE 11
   3.3 CONVENTIONS 12
   3.4 CONCEPT 12
   3.5 POSSIBLE IMPLEMENTATION 13
   3.6 EXAMPLES IN DATASOURCES PRODUCTION 13
      3.6.1 OJS 13
      3.6.2 DSPACE 14
      3.6.3 EXAMPLE WITH SIGNPOSTING: 14

4. TECHNICAL REFERENCES 15
Disclaimer

This document contains description of the OpenAIRE-Advance project findings, work and products. In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the OpenAIRE-Advance consortium and can in no way be taken to reflect the views of the European Union.

OpenAIRE-Advance is a project funded by the European Union (Grant Agreement No 777541).
Publishable Summary

The aim of this report is to present the rationale for the adoption of two “Next Generation Repository” recommended technologies and protocols: ResourceSync and Signposting. Both of these technologies contribute to improving the discovery of repository content in repositories and contribute to several of the behaviours described in the COAR Next Generation Repository report.
1 | INTRODUCTION

Repositories forming a distributed global knowledge network have the potential to promote the transformation of the scholarly communication ecosystem. One of the barriers is a lack of web-based integration with other innovative scholarly services. There is a clear need for repository platforms to adopt modern web-technologies and protocols that will allow them to better interact with more innovative and sophisticated scholarly-networked tools and services.

This report is a first outcome of the OpenAIRE Advance project task number 6.2 on the Next generation repositories (NGR) that aims to set up the specifications of the NGR technologies and protocols and establish the priorities for the implementation of technologies and services within the OpenAIRE infrastructure.

This report outlines the rationale for the adoption of two “Next Generation Repository” recommended technologies and protocols: ResourceSync and Signposting. Both of these technologies contribute to improving the discovery of repository content in repositories and contribute to several of the behaviours described in the COAR Next Generation Repository report.
RESOURCESYNC

ResourceSync is a mechanism for large-scale synchronisation of web resources. NGR promotes the use of ResourceSync as a modern way to allow aggregation services to 'harvest' metadata and content from repositories. Although ResourceSync can be used in a range of ways, NGR recommends that ResourceSync is used to expose changes in repository content, so that the repository's content (and/or metadata) may be reliably synchronised with other services (including aggregators).

What is the problem? Currently, aggregators use OAI-PMH to harvest metadata from repositories. While OAI-PMH is very well-supported by repositories, it has some limitations which make it unsuitable for modern, large scale harvesting. Firstly, it supports the harvesting of metadata only. Most modern aggregation services need to access, harvest and aggregate a repository's content - as well as its metadata - but OAI-PMH does little to support this. Secondly, OAI-PMH does not enforce a reliable way of conveying information about items which may have been removed from a repository. This means that, over time, aggregations become 'out-of-sync' with the repositories they harvest. With OAI-PMH, eventually the aggregation service is forced to start afresh and re-harvest everything from a given repository.

What is ResourceSync? ResourceSync is a specification based on Sitemaps that can be implemented into repositories allowing third-party systems to remain in sync with the resources in their repository as they evolve, i.e. are created, updated, deleted. ResourceSync automates the replication and updating process ensuring that discovery services reflect current content in repositories.

The ResourceSync Framework Specification describes a synchronization framework for the web consisting of various capabilities that allow third-party systems to remain synchronized with a server’s evolving resources. More precisely the ResourceSync Framework describes the communication between Source and Destination aimed at synchronizing one or more resources. Communication utilizes http and an extension on the site map protocol, an xml-based format for expressing metadata, relevant for synchronization.

2.1 Use Cases

Use Case 1 - Metadata Harvesting: A service uses ResourceSync to harvest and synchronise metadata about content held in many repositories.

This is a classic repository use-case, and is the one currently supported by OAI-PMH which is supported by almost all repositories. In this use-case, only the metadata is synchronised, and users of the service are directed back to the 'source' repository if they want to retrieve the actual content described in the synchronised metadata.
The repository and the service may each decide which metadata is harvested. While the typical case might be for the service to harvest all available metadata from a given repository it might, instead, be selective in its harvesting.

This allows the creation of 'discovery' services which aggregate - and then provide search functions across - metadata from many repositories.

- Repositories are registered with the service.
- Repositories expose their metadata through a Resource List, which conveys the URIs of resources available for synchronization. The Resource List is expressed as a Sitemap and provides the base level of interoperability for ResourceSync.
- Repositories communicate information to the service about any changes to its resources by publishing a Change List, which enumerates resource changes, along with the nature of the change (create, update, or delete) and metadata describing the new state of the resource.
- The service can recurrently obtain a Change List from the repository, inspect the listed changes to discover those it has already acted upon, and process the remaining ones.

**Use Case 2 - Content Harvesting:** An 'aggregation' service uses ResourceSync to harvest and synchronise content from many repositories. This is the more sophisticated version of use-case 1, where the service harvests not only the metadata but also the content from many repositories (something which is not supported by OAI-PMH).

The repository and the service may each decide which content is harvested. It is quite possible that a service might be only interested in certain types or formats of content, or it might be interested in a subset of a repository's content from, for example, a particular domain. The ResourceSync protocol supports selective harvesting.

While the typical case might be for the service to harvest all available content from a given repository it might, instead, be selective in its harvesting.

This allows the creation of more sophisticated discovery services which, for example, utilise text and data-mining techniques to extract more information about the content.

- Repositories are registered with the service.
- Repositories expose their resources through a Resource List, which conveys the URIs of resources available for synchronization. The Resource List is expressed as a Sitemap and provides the base level of interoperability for ResourceSync.
- Repositories communicate information to the aggregator about any changes to those resources by publishing a Change List, which enumerates resource changes, along with the nature of the change (create, update, or delete) and metadata describing the new state of the resource.
The aggregator can recurrently obtain a Change List from the repository, inspecting this to discover new changes that the aggregator has not yet processed, and then applying those changes to the aggregator's copy of the resource.

2.2 Implementing ResourceSync in your repository

- There have been Plug-ins developed for both [DSpace 5 and 6 Module](#) for ResourceSync and DSpace 7 will have ResourceSync built into the main codebase.
- [Hyrax (Samvera)](#) supports ResourceSync, providing resource lists and change lists
- EPrints is working on an [Plug-in](#)

2.3 Service providers using ResourceSync

- The [CORE](#) aggregator has been piloting and working to optimize ResourceSync across a number of use cases
- The [Digital Public Library of America (DPLA)](#) is using [ResourceSync](#) to harvest from some repositories

More information is available on the NISO website [here](#), and the user stories related to ResourceSync are available on the COAR NGR website ([ResourceSync](#)). Also see a 2018 [presentation](#) by Martin Klein from Los Alamos National Laboratory.
3 | SIGNPOSTING

Signposting is an approach to inform machine agents about the nature of the resources that are linked from the resource they currently interact with. It uses typed links (in the HTTP Link header, the HTML element, or the rs:ln ResourceSync element) to reveal patterns that occur repeatedly in scholarly portals. Signposting can be used to support automatic discovery of a variety of resources that pertain to a scholarly object, including a bibliographic description, a persistent identifier, a license, authors, or various resources that are part of the object. [http://signposting.org](http://signposting.org)

This part introduces the discussion of possible use cases to be included in the context of OpenAIRE. These use cases will support the technical team to test and implement services based on the approach of Signposting. The main objective is to build the demonstrator for signposting.

Patterns:
The patterns defined by Signposting are:

- identifier
- bibliographic metadata
- author
- publication boundary

Constraints

- not all patterns must/need to be implemented and supported
- in some cases, the pattern is not implemented with all supported links

Open questions

- Define a minimum set of data to be signposting/openaire compliant
- Benefits for different stakeholders
- How can OpenAIRE benefit from data providers supporting signposting?

3.1 Use Cases

UC1 - Discovery through Navigation

The Signposting approach allow the discovery and navigation through the different nodes and links.

one starting point could be: [http://ngr.coar-repositories.org/behaviour/discovery-through-navigation/](http://ngr.coar-repositories.org/behaviour/discovery-through-navigation/)
- Online Navigation / browsing LOD - Linked Open Data

UC2 – Enrichment

UC2.1 - Enrichment with Identifiers

- Links and connections of “enhanced publications”

UC2.2 - Enrichment with Metadata

- enrich existing info
by following URLs
- by evaluating URLs, e.g.
  - identifiers (DOI, URN, ORCID, ...) as part/postfix of URL strings
  - access rights deduced from license link strings
- by following resolver links plus PIDs (e.g. to obtain metadata of cited publications)

○ identify fulltext

**UC2.3 - Enrichment with Usage Metrics**
- usage metrics

**UC3 - Update**
- get updated/modified data?

**UC4 - Harvest**
- collecting all entries possible
- Process data in different steps:
  - 1 - get headers
  - 2 - process other links (items, metadata, authors...)
  - 3 - relate information for the end user

**UC5 - Simple Web Service Concept**
Think about Signposting for simple/micro service.

### 3.1.1 Uses Cases by Service

How can we get and reuse this data?
- International Harvester
  - Harvest metadata through signposting or enrich existing metadata
- National Harvester
- Institution - to get all publications from authors from that institution?
- Repository

### 3.2 Signposting Patterns

#### 3.2.1 Author - [https://signposting.org/author/](https://signposting.org/author/)

Use **Typed Links** in **HTTP Link** headers to help machines find the authors of a publication.

The following relation type is used in the Author pattern:

To link from an **Identifying HTTP URI**, a **Repository Stable HTTP URI**, or an **Entry Page** to a URI identifying the author: **author**. Author identifiers that are conveyed on links with the "author" relation type should be HTTP URIs chosen from namespaces that are widely supported by the community, such as **ORCID**, **VIAF**, and **ISNI**.
3.2.2 Bibliographic Metadata

https://signposting.org/bibliographic_metadata/

Use Typed Links in HTTP Link headers to help machines find the bibliographic metadata that describe a publication.

Several metadata resources may exist, each expressed according to a different format. Each such metadata resource should be published at a URI that supports HTTP GET. Currently, many portals exist where this is not the case. Also, the format of each metadata resource should be expressed unambiguously in its HTTP response header. When it comes to bibliographic metadata resources, there is currently little uniformity out there. We propose some conventions to start addressing that problem.

The following relation types are used in the Bibliographic Metadata pattern:

- To link to and from bibliographic metadata, respectively: describedby, describes

In addition, the following attributes should be used on Typed Links:

- To express the media type of the linked resource: type
- To provide further details about the media type of the linked resource, for example, when several linked resources have the same media type: profile

3.2.3 Identifier

https://signposting.org/identifier/

Use Typed Links in HTTP Link headers to convey the preferred Identifying HTTP URI of a resource to machines. The Identifying HTTP URI conveyed in this manner serves as a replacement of the URI of the resource that conveys the link and is intended for persistent referencing.

The following relation type is used in the Identifier pattern:

- To convey the preferred Identifying HTTP URI of a resource: cite-as

3.2.4 Publication Boundary

https://signposting.org/publication_boundary/

Use Typed Links in HTTP Link headers to help machines find the resources that make up a publication.

The following relation types are used in the Publication Boundary pattern:

- To connect the resources that make up a publication: item, collection
- To express the media type of the linked resource: type

3.2.5 Resource Type

https://signposting.org/resource_type/
Use **Typed Links** in **HTTP Link headers** to convey the semantic type of a resource to machines.

The following relation type is used in the Resource Type pattern:

- To express the semantic type of an **Identifying HTTP URI**, a **Repository Stable HTTP URI** or an **Author Identifying HTTP URI**: `type`. To align with web best practice, semantic types that are conveyed on links with the "type" relation type should be chosen from the **schema.org** ontology. But, additionally, HTTP URIs from other ontologies can be chosen if more expressiveness is required.

### 3.3 Conventions

[http://signposting.org/conventions/](http://signposting.org/conventions/)

Signposting proposes a really simple approach, based on Typed Links conveyed in HTTP Link headers, to clarify patterns that occur repeatedly in scholarly portals. Typed Links are used to help machines answer questions like "What is the DOI of this PDF publication", "Where to find the publication resources from the landing page?", "Where to find the BibTeX metadata that describes the publication", "What is the ORCID of the author of this publication".

### 3.4 Concept
3.5 Possible Implementation

- DSpace - https://jira.duraspace.org/browse/DS-3589
- OJS - https://github.com/4Science/signposting
- Search Portal - OpenAIRE / RCAAP / La Referência

3.6 Examples in datasources production

There are already some datasources that have signposting patterns implemented. The following examples of early adopters are available through the website: https://signposting.org/adopters/

- DANS
- DataCite Search
- DSpace-CRIS
- Open Journal System
- Pangaea
- UCD Digital Library

There is also the LibreCat demonstrator implementation available here: https://github.com/LibreCat/LibreCat/wiki/Signposting.

3.6.1 OJS

https://www.annalsofgeophysics.eu/index.php/annals


HTTP/1.1 200 OK
Date: Thu, 18 Oct 2018 10:13:17 GMT
Server: Apache
Set-Cookie: OJSSID=39ubmea41ub7a55g20go5som74; path=/; domain=www.annalsofgeophysics.eu
Link: <https://www.annalsofgeophysics.eu/index.php/annals/sp-citation/bibtex/7356>; rel="describedby"; type="application/x-bibtex",
<https://www.annalsofgeophysics.eu/index.php/annals/sp-citation/endNote/7356>; rel="describedby"; type="application/x-endnote-refer",
<https://www.annalsofgeophysics.eu/index.php/annals/sp-citation/proCite/7356>; rel="describedby"; type="application/x-research-info-systems",
<https://www.annalsofgeophysics.eu/index.php/annals/sp-citation/refWorks/7356>; rel="describedby"; type="application/x-refworks",
<https://www.annalsofgeophysics.eu/index.php/annals/sp-citation/refMan/7356>; rel="describedby"; type="application/x-research-info-systems",
<https://doi.org/10.4401/ag-7356>; rel="describedby"; type="application/vnd.citationstyles.csl+json",<https://doi.org/10.4401/ag-7356>; rel="cite-as",<https://doi.org/10.4401/ag-7356>
3.6.2 DSpace

**item example**

HTTP/1.1 303 See Other  
Server: Apache-Coyote/1.1  
Location: http://repositorium.sdum.uminho.pt/handle/1822/54996  
Expires: Thu, 18 Oct 2018 09:48:26 GMT  
Content-Type: text/html; charset=utf-8  
Content-Length: 189  
Date: Thu, 18 Oct 2018 09:46:45 GMT  
curl -I http://hdl.handle.net/1822/54996

3.6.3 Example with Signposting:

HTTP/1.1 303 See Other  
Server: Apache-Coyote/1.1  
Location: http://repositorium.sdum.uminho.pt/handle/1822/54996  
Expires: Thu, 18 Oct 2018 09:48:26 GMT  
Content-Type: text/html; charset=utf-8  
Content-Length: 189  
Date: Thu, 18 Oct 2018 09:46:45 GMT

<!-- New code -->
<!-- link to citation -->
Link: <http://hdl.handle.net/1822/54996> ; rel="cite-as" ,
<!-- link to item file -->
Link:  
<http://repositorium.sdum.uminho.pt/bitstream/1822/54996/2/OR2018_poster_OpenAIRE.pdf>  
; rel="item" ; type="application/pdf" ,
<!-- Author Identifiers →
Link: <https://orcid.org/0000-0002-8588-4196> ; rel="author" ,
  <https://orcid.org/0000-0002-0458-1004> ; rel="author" ,
<!-- Relation is described by →
Link: <http...> ; rel="describedby" ; type="application/x-bibtex" ,
4 | TECHNICAL REFERENCES


