Towards a global knowledge commons

Next Generation Repositories

February 7, 2017 – draft for public comment

http://coar-repositories.org

Contents

Next Generation Repositories

February 7, 2017 – draft for public comment

Public comments are open from February 7 – March 3, 2017

In April 2016, the Confederation of Open Access Repositories (COAR) launched a working group to help identify new functionalities and technologies for repositories and develop a road map for their adoption. For the past several months, the group has been working to define a vision for repositories and sketch out the priority user stories and scenarios that will help guide the development of new functionalities.

The vision is to position repositories as the foundation for a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value added services will be deployed, thereby transforming the system, making it more research-centric, open to and supportive of innovation, while also collectively managed by the scholarly community.

Underlying this vision is the idea that a distributed network of repositories can and should be a powerful tool to promote the transformation of the scholarly communication ecosystem. In this context, repositories will provide access to published articles as well as a broad range of artifacts beyond traditional publications such as datasets, pre-prints, working papers, images, software, and so on.

The working group presents 12 user stories that outline priority functionalities for repositories.

Please contribute your ideas and opinions using the commenting function of the website!

All comments will be reviewed and considered by the working group

Once input has been incorporated and the user stories are finalized, the working group will begin to outline recommended technologies and architectures to support the adoption of functionalities into repository platforms.

COAR Next Generation Repositories Working Group

Eloy Rodrigues, chair (COAR, Portugal)

Andrea Bollini (4Science, Italy)

Alberto Cabezas (LA Referencia, Chile)

Donatella Castelli (OpenAIRE/CNR, Italy)

Les Carr (Southampton University, UK)

Leslie Chan (University of Toronto at Scarborough, Canada)

Chuck Humphrey (Portage, Canada)

Rick Johnson (SHARE/University of Notre Dame, US)

Petr Knoth (Jisc and Open University/CORE, UK)

Paolo Manghi (CNR, Italy)

Lazarus Matizirofa (NRF, South Africa)

Pandelis Perakakis (Open Scholar, Spain)

Jochen Schirrwagen (University of Bielefeld, Germany)

Daisy Selematsela (NRF, South Africa)

Kathleen Shearer (COAR, Canada)

Tim Smith (CERN, Switzerland)

Herbert Van de Sompel (Los Alamos National Laboratory, US)

Paul Walk (EDINA, UK)

David Wilcox (Duraspace/Fedora, Canada)

Kazu Yamaji (National Institute of Informatics, Japan)

Rationale

In April 2016, COAR launched a working group to help identify new functionalities and technologies for repositories and develop a road map for their adoption. The aim of this activity is to develop a global network of repositories that allows frictionless access to open content and encourages the creation of cross-repository added-value services. The approach chosen to achieve this goal consists of two threads:

1. The first aims to increase the exposure by repositories of uniform behaviors that can be used by machine agents to fuel novel scholarly applications that reach beyond the scope of a single repository and that enable to smoothly embed repository content into mainstream web environment. Currently, OAI-PMH is the only behavior that is uniformly exposed by most repositories. While OAI-PMH undoubtedly has merits, its focus on metadata, its pull-based paradigm, and its technological roots that date back to the web of the nineties put it at odds with the realities of current scholarly practice and web technologies.

1. The second thread consists of integrating with existing scholarly infrastructures, specifically those aimed at identification, as a means to solidly embed repositories in the overall scholarly communication landscape. It is time to reimagine the potential of repositories and to mobilize the global repository community towards selecting and implementing appropriate technologies that can help realize that potential.

These threads will help to re-position the global repository network and support the transformation of the scholarly communication system into one that is collective, open and managed in a distributed manner.

Vision and Objectives

Vision

To position repositories as the foundation for a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value added services will be deployed, thereby transforming the system, making it more research-centric, open to and supportive of innovation, while also collectively managed by the scholarly community.

Objectives

- To achieve a level of cross-repository interoperability by exposing uniform behaviours across repositories that leverage web-friendly technologies and architectures, and by integrating with existing global scholarly infrastructures specifically those aimed at identification of e.g. contributions, research data, contributors, institutions, funders, projects.

- To encourage the emergence of added-value services that use these uniform behaviours to support discovery, access, annotating, real-time curating, sharing, quality assessment, content transfer, analytics, provenance tracing, etc.

- To help transform the scholarly communication system by emphasizing the benefits of collective, open and distributed management, open content, uniform behaviours, real-time dissemination, and collective innovation.

Principles and Design Assumptions

Principles

Distribution of control: Distribution of the control of *scholarly resources* (pre-prints, post-prints, research data, supporting software, etc.) and *scholarly infrastructures* is an important principle which underpins this work. Without this, a small number of actors can gain too much control and can establish a quasi-monopolistic position – as has happened in the scholarly publishing industry. Note that this does not require the distribution of *systems*, although whatever solution we develop must certainly be able to thrive in and support such an architecture.

Inclusiveness: It is acknowledged that different institutions and regions have unique and particular needs (e.g diverse language, policies and priorities) and this will be reflected in the network.

Public good: The technologies, architectures and protocols adopted in the context of the global network for repositories will be available to everyone, wherever possible using global standards.

Intelligent openness: Scholarly resources, wherever possible, will be made openly available in order increase their value and maximize their benefit for scholarship and society.

Design Assumptions

Focus on the resources themselves, not just associated metadata: For historical reasons, technical solutions have focused on metadata that describes scholarly resources instead of on the resources themselves. By considering both the scholarly resource and its metadata as web resources identified by distinct URIs, they can be treated on equal footing and can be appropriately interlinked.

Pragmatism: Given the choice, we tend to favour the simpler approach. Where possible, we choose technologies, solutions and paradigms which are already widely deployed. In practical terms, this means that we favour using standard Web technologies wherever possible.

Evolution, not revolution: We prefer to *evolve* solutions, adjusting existing software and systems where possible, to better exploit the ubiquitous Web environment within which they are situated.

Convention over configuration: Our preference is to adopt widely recognised conventions and encouraging everyone to use these where possible, rather than accommodating richer, more complex and varied approaches. As a corollary to this, we favour standardising only that which needs standardising and keeping constraints to a minimum, so that those implementing our systems can readily understand the constraints under which they must operate.

Engage with users where they are: Instead of always asking users to leave their environment and engage with one of our systems, integrate tools into the environments and systems where they are already engaged.

User Stories

The next generation repository welcomes both human and machine users. As such, the use cases presented below address both of these scenarios. Ideally, users could interact with different repositories more or less in a common manner. Although humans are flexible and can handle diverse user interfaces, uniformity is crucial, when welcoming machine users. Uniform interfaces enable interaction with a plethora of distributed repositories in a standard way, for example, to batch discover content or metadata, or to navigate from one repository resource to another. They support machine tasks such as creating a cross-repository search engine, or collecting resources for preservation or data mining purposes. They also allow the creation of more intelligent tools that improve the experience for human users.

1. Discovering metadata that describes a scholarly resource

As a human or machine user, I want to easily and uniformly identify the metadata in a repository record, so that I can ascertain the relevance of the resource. Most repositories provide links to bibliographic information that describes a scholarly resource. Typically, these links are provided in the landing page and are discriminated by the use of link anchors that identify a metadata schema or citation format such as “bibtex”, “RIS”, “DC”, etc. This allows a researcher to easily select the desired metadata, if needed. But tools such as reference managers or crawlers that are on a digital preservation or data mining mission cannot easily or uniformly find their way to that metadata. They need to resort to repository-specific heuristics when trying to accomplish their task. Also, when these tools land on resources other than the landing page – say the PDF or the dataset – they are at a loss as to where to find the metadata. Using typed HTTP links with appropriate link types and format indicators on web resources that make up a scholarly object will enable both humans and machines to discover the metadata and accomplish their tasks.

2. Discovering the identifier of a scholarly resource

Web reference managers, annotation tools, or crawlers that encounter a landing page or any other web resource that is part of a scholarly object need to easily identify the associated persistent HTTP URI for the resource, so that they can retrieve it. Many repositories assign persistent identifiers to the scholarly resources they host. Since repositories reside on the web, the persistent identifier is expressed as a HTTP URI. The persistent HTTP URI is in most cases distinct from the URI of the landing page. As a matter of fact, it typically redirects to the landing page. Also, the actual content – say the PDF or the dataset – resides at yet another HTTP URI. As a result, in many cases, authors refer to resources by means of their landing page URI or the URI of actual content, even though the landing pages of some repositories indicates – in a human-readable manner – that the persistent HTTP URI should be used for referencing. When reference managers, annotation tools, or crawlers happen upon a landing page or any other web resource that is part of a scholarly object, they are unable to identify the associated persistent HTTP URI. This is rather detrimental as the investment that is made in trying to achieve persistence goes to waste. This problem can be addressed by using typed HTTP links with an appropriate link type to point from web resources that are part of a scholarly object to their persistent HTTP URI. This allows tools – potentially even the browser bookmarking tool – to auto-discover the identifier. Authors no longer need to bother to copy/paste the identifier from the landing page. And the persistence intended by these identifiers is achieved.

3. Discovering usage rights

As a machine or human user, I need to easily and uniformly identify the licensing and re-use conditions of a scholarly resource, so that I know what I am allowed to do with it. Ideally, scholarly objects would be available without constraints on how they can be used. The reality is different, however, and in many cases limitations do apply. If so, these limitations should be clearly indicated for each web resource that is part of a scholarly object and they should be discoverable by both human and machine users. For humans, this can, for example be achieved by embedding easily recognizable logos that convey the license that applies. For machines, this can be achieved by using appropriately typed HTTP links that point at the URI of the license that applies. Once licenses are exposed in this manner, tools such as reference managers can convey this information to humans that use the tool and store it in their database. Crawlers that are on a digital preservation or data mining mission can act according to the constraints imposed by the license when deciding whether to collect and how to further handle a resource. The use of common licenses, such as those provided by the Creative Commons, makes it easy for both humans and machines to readily understand which constraints apply.

4. Recognizing the user

As a user, I want my repository to recognize me and other users so that I can be connected with other users who I know, leave comments and be informed of content that is of interest to me. It is very common for blogging platforms to allow readers to comment on postings by authors that use the platform. When these readers are required to identify themselves to comment and this can lead to constructive conversations and the creation or reinforcement of social connections. Similarly, repositories can allow readers to comment, annotate, or peer-review scholarly resources authored by the scholars that deposit to the repository. Identification of readers can be done by means of a permanent ID if they are researchers or by another web identity such as the popular mainstream social networks. This can help foster a new level of scholarly interactions. But, additionally, the ability to uniformly identify oneself in repositories, worldwide, adds a global dimension to repositories that have thus far been largely isolated from each other.

5. Commenting, annotating, and peer-review

As a user, I want to be able to comment or review the work of my colleagues and have those reviews (and reviewers) publicly available to all readers, so that the quality of these resources are assessed by others. Related to the previous user story, repositories can increase their value by supporting commenting, annotating and peer review activities as functional layers on top of their collective content. If repositories were able to support assessment and peer review, they could begin to reposition themselves at the centre of scholarly communication. To that end, repositories will need to support services that allow researchers to comment and annotate the papers deposited in the repositories, in an open, sharable and interoperable way, so that discussions and collaborative work can be promoted. The types of functionalities needed for this are described below:

As an ***author*** I want to:

1. Invite any expert peer on the subject to review my manuscript, in order to improve its quality
2. Have a fluid communication with the reviewer(s) during the review process (e.g., be able to exchange quick comments to clarify issues, chat, etc).
3. Be able to update my work based on reviewer comments/suggestions and upload it as new version.

4. Be able to publicly acknowledge reviewers for helping improve my work (e.g., through badges, ratings, etc).
5. Have access to a list of active reviewers in my discipline.

As a ***reviewer*** I want to:

1. Be able to review any document in the archive, either after invitation or following my own initiative
2. Have a fluid communication with the author(s)
3. Receive recognition for my review by the authors and the community (e.g., acknowledgment from authors, ratings from community, etc.)

As a ***reader*** I want to:

1. Have access to the full text submitted by each reviewer for a given manuscript version
2. Have access to the name of the reviewers and possible conflicts of interest (e.g., recent collaboration with the author).
3. Have access to easily interpretable quantitative data about each manuscript informing about its validity and perceived importance.
4. Be able to filter and sort works in the archive (e.g., based on number of reviews, review scores, etc.).
5. Be able to suggest reviewers for specific manuscripts/works.
6. Have quick access to all reviews made by a specific reviewers along with quantitative information about this reviewer's performance

6. Automated recommender systems for repositories

As a user, I want to receive recommendations about content that is of potential interest to me and related to my work, so I increase my knowledge in my field. Recommendation systems can significantly help users who are navigating large-scale research collections and datasets. Recommender systems are believed to be one of the key functionalities making academic social network sites popular. As a repository user, I want to be able to discover new research outputs related to my interest, both pro-actively when browsing as well as in the form of notifications, regardless of the place in which they are stored. Equally, I want to be able to discover and identify important people, relevant scientific methods, conference/journal/meetup venues, funding opportunities, etc. in the research field I am interested in. Recommendation systems today are typically based on related content or user-interaction based. The most successful systems use a combination of the two. Repositories currently only make it possible to build content-based recommender systems.

To enable the creation of state-of-the-art recommender systems for repositories, they will need to:

- Offer a machine interface providing access to anonymised user-interaction logs, in the form of a triple <user, activity, time>, where user is either an anonymous session or an anonymized (e.g. hashed) user id and where activity is, for example, a download or view event.
- Enable a secure cross-repository user profile (for personalised recommendations and notification) holding the user's interests which can be user specified or derived based on previous interactions with any of the repository systems in the global network. Such profiling can also be based partly on an existing service.

7. Providing a social notification feed

As a repository user, I want to have access to a global, cross-repository social feed so that I am informed about activities in which I have registered an active interest. For example, one of my social media contacts added a document, someone commented on a paper in a feed I was subscribed to, an open review has been provided on a paper I have read, a new dataset has been attached to a paper I am watching, a paper has been published based on a dataset I have used, etc. In order to be able to support this functionality, repositories should be able to actively and in real-time push activity events (document changes, additions, comments, new peer-reviews, etc) to "global" (but possibly distributed) interaction hubs; and repositories (and possibly directly users) should be able to consume in real-time activity event logs from the interaction hubs. Content curated and available through repositories should be embeddable and taggable within communication platforms, broadcasted through common social media channels. Subsequently, content from related activity within these platforms should be curatable as well.

8. Resource syncing and notification

As a repository manager, I want my repository to be automatically notified about new or modified relevant objects and metadata, so that I can have a more complete and accurate collection. I also want other, remote systems to be notified of changes made to my collection to ensure that records are standardized across various locations. To fulfill this use case, the repository will need to be notified from other services about new and related content via a hub through which it has subscribed 'interest' in notifications. To support this, a repository will need to expose all changes to its contents (new objects, or new/enhanced version of existing objects, or new/ edited/ enhanced metadata records related with those objects) to all interested repositories or other information systems. Simultaneously, the repository will be notified of changes of content in other repositories/information systems that are relevant for its context (based on identifiers – authors, objects and institutions).

9. Data mining

As a human or machine user, I want to be able to mine text or data in the collective content of repositories to discover new relationships and make new discoveries. Repositories should enable the data mining of their content by allowing third parties to effectively access and transfer full text objects. The data mining activity is carried out outside of the repository network in a data miner's infrastructure of choice. Such third parties include aggregators, which need to be able to effectively and in a timely manner transfer the available data from repositories into another location. Furthermore, they need to be able to effectively and in a timely manner synchronise their own dataset with the repositories from which they are harvesting full text, and they must be able to track all changes to the underlying datasets, such as updates, deletions and additions. To provide additional support for the reproducibility of data mining experiments, repositories should expose a log of all changes to the underlying data. The data offered via the repository machine interfaces must include the metadata and full text content, as well as a log of all changes. In addition, repositories should enable the sharing of user interaction data, such as clicks (article level CTR), co-downloads or comments, to enable the development, deployment and evaluation of innovative value-added global services over repositories.

10. Supporting researchers' workflows

As a researcher, I want to easily deposit my paper with as little effort as possible into the repository platform, so that my paper is available to comply with policy requirements. In this scenario, the repository needs to be interoperable or integrated with other tools that authors are using offline and collaborative online tools such as microsoft word or google docs, as well as with journal submission systems such as the OJS platform. Additionally the repository should automatically recognize and autofill the metadata from my paper, dataset or other objects including author, title, date and so on into the submission form.

11. Comparing usage

As a user, I want to know how many times a paper has been downloaded and read by others according to a standard metric, to be able to assess its impact. Collecting metrics is relevant to optimise, operate, and enhance the repository and to demonstrate the value of the repository to authors and other stakeholders. As an author, I want to know how often my paper or dataset is being read and cited, and to be able to compare that with other papers of my peers so that I have an objective, standardized way of assessing the impact of my work. Methodologies for measuring usage must be standardized across repositories and repository platforms. The measures also need to be trusted by the community as accurate and reliable so that we can begin to incorporate them into discussions about the benefits of open access.

12. Preservation

As a scholar, I want my research outputs to be available over the long term and remain as a permanent part of the scholarly record. I also want to know that my article will be recoverable in the event a repository loses its copy of my work. I may also be interested in searching archival holding.Open access means not just that you can have access to things today, but also into the future. We can envision preservation services that will support repository operations within a network. Not every repository needs to run its own preservation processing stack, but rather we need common standards, protocols and interoperability that will enable us to build these services for repositories in a collective way. Additionally it is necessary to preserve the complex interconnection of resources, which involves preservation activities at various levels including the resource, metadata and information graph. Furthermore, through enhanced clients and embedding new technology in information creation and communication platforms, capture and preserve content creation in real-time.

What have we missed?

Please add your own user story, if you think we missed something important!

Added by the Next Generation Repsoitories Working Group after publication

Batch-based content discovery

A user wants to discover repository materials via aggregators such as Google Scholar, CORE, … And in order to make that happen, repositories need to implement widely supported batch content discovery mechanisms such as Sitemaps. And, if in addition metadata about content needs to be discovered, ResourceSync (based on Sitemaps) can be used because it allows lining to related resources, including metadata describing resources.