Next Generation Repositories

Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group

November 28, 2017

Northern lights, Norway

#nextgenrepositories @COAR office@coar-repositories.org

In April 2016, the Confederation of Open Access Repositories (COAR) launched the Next Generation Repository Working Group to identify new functionalities and technologies for repositories. In this report, we are pleased to present the results of the work of this group, including recommendations for the adoption of new technologies, standards, and protocols that will help repositories become more integrated into the web environment and enable them to play a larger role in the scholarly communication ecosystem.

The current system for disseminating research, which is dominated by commercial publishers, is far from ideal. In an economic sense, prices for both subscriptions and APCs are over-inflated and will likely continue to rise at unacceptable rates. Additionally, there are significant inequalities in the international publishing system both in terms of access and participation. The incentives built into the system, which oblige researchers to publish in traditional publishing venues, perpetuate these problems and greatly stifle our ability to evolve and innovate.

At COAR, we believe the globally distributed network of more than 3000 repositories can be leveraged to create a more sustainable and innovative system for sharing and building on the results of research. Collectively, repositories can provide a comprehensive view of the research of the whole world, while also enabling each scholar and institution to participate in the global network of scientific and scholarly enquiry. Building additional services such as standardized usage metrics, peer review and social networking on top of a trusted global network of repositories has the potential to offer a viable alternative.

The vision underlying the work of Next Generation Repositories 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."

An important component of this vision is that repositories will provide access to a wide variety of research outputs, creating the conditions whereby a greater diversity of contributions to the scholarly record will be accessible, and also formally recognized in research assessment processes.

Our vision is aligned with others, such as MIT's Future of Libraries Report[1] and Lorcan Dempsey's notion of the "inside-out" library[2], that are defining a new role of libraries in the 21st century. This future involves a shift away from libraries purchasing content for their local users, towards libraries curating and sharing with the rest of the world the research outputs produced at their institution. COAR's mission is to ensure that, as libraries and research organizations invest in and enhance their local services, they adopt common

[1] https://future-of-libraries.mit.edu/
[2] https://www.liberquarterly.eu/articles/10.18352/lq.10170/

standards and functionalities that will allow them to participate in the global network. We very much hope that the recommendations provided in this report will contribute to the transition towards this new role for repositories and libraries.

This was a truly collaborative effort. We would like to sincerely thank the members of the Next Generation Repositories Working Group for their generous contributions and significant efforts towards this undertaking. They have brought a breadth and depth of expertise, without which we would not have been able to accomplish this work. We are very grateful!

Eloy Rodrigues, COAR Chairman and Kathleen Shearer, COAR Executive Director

Executive Summary

The widespread deployment of repository systems in higher education and research institutions provides the foundation for a distributed, globally networked infrastructure for scholarly communication. However, repository platforms are still using technologies and protocols designed almost twenty years ago, before the boom of the Web and the dominance of Google, social networking, semantic web and ubiquitous mobile devices. This is, in large part, why repositories have not fully realized their potential and function mainly as passive recipients of the final versions of their users' conventionally published research outputs. In order to leverage the value of the repository network, we need to equip it with a wider array of roles and functionalities, which can be enabled through new levels of web-centric interoperability.

In April 2016, COAR launched the Next Generation Repositories Working Group to identify the core functionalities for the next generation of repositories, as well as the architectures and technologies required to implement them. This report presents the results of work by this group over the last 1.5 years.

> *"Our 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."*

The next generation repository...

- manages and provides access to a wide diversity of resources, including published articles, pre-prints, datasets, working papers, images, software, and so on.
- is resource-centric, making resources the focus of its services and infrastructure
- is a networked repository. Cross-repository connections are established by introducing bi-directional links as a result of an interaction between resources in different repositories, or by overlay services that consume activity metadata exposed by repositories
- is machine-friendly, enabling the development of a wider range of global repository services, with less development effort
- is active and supports versioning, commenting, updating and linking across resources

The Next Generation Repositories Working Group has explicitly focused on the generic technologies required by all repositories to support the adoption of common behaviours. However, we also recognize that there are other technologies and standards that may be useful for specific content types or disciplinary communities.

This report describes 11 new behaviours, as well as the technologies, standards and protocols that will facilitate the development of new services on top of the collective network, including social networking, peer review, notifications, and usage assessment.

1. Exposing Identifiers
2. Declaring Licenses at a Resource Level
3. Discovery through Navigation
4. Interacting with Resources (Annotation, Commentary and Review)
5. Resource Transfer
6. Batch Discovery
7. Collecting and Exposing Activities
8. Identification of Users
9. Authentication of Users
10. Exposing Standardized Usage Metrics
11. Preserving Resources

The behaviours and technologies in this report are a snapshot of the current status of technology, standards and protocols available, but we are aware that technologies will continue to evolve. To that end, we will soon be publishing the behaviours and technologies in a GitHub repository to support updates, as well as enabling greater input and engagement with the broader community as technologies evolve or new technologies come onto the scene.

In conclusion, the distributed network of repositories can and should be a powerful tool to promote the transformation of the scholarly communication ecosystem, making it more research-centric, innovative, while also managed by the scholarly community. However, this vision rests on the notion that repositories behave (or function) in common ways, and interact with external services in the same manner. As such, it is important that the technologies, standards and protocols defined here are widely accepted and adopted by repositories around the world.

Next Generation Repositories Working Group

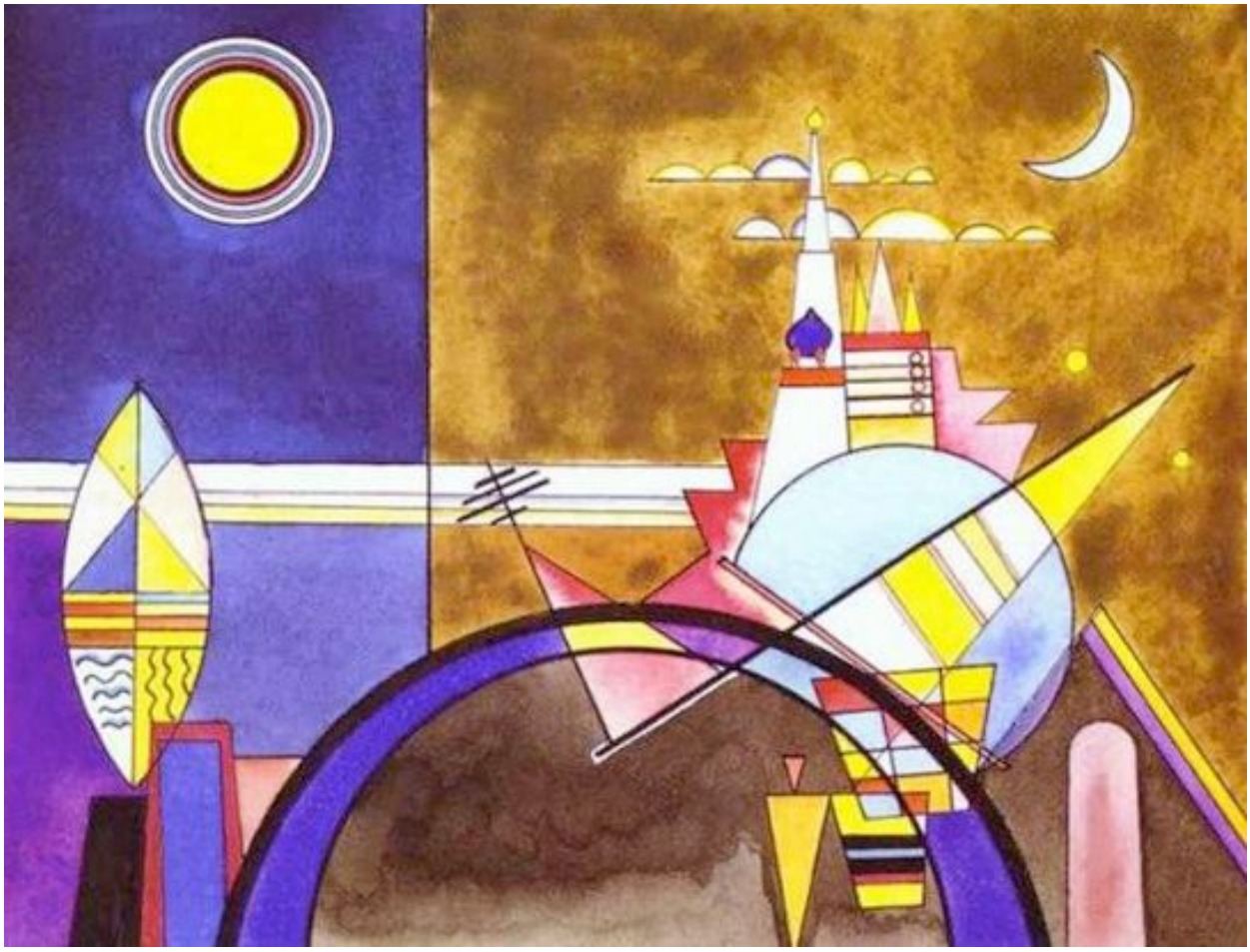
Eloy Rodrigues, chair (COAR, Portugal)
Andrea Bollini (4Science, Italy)
Alberto Cabezas (LA Referencia, Chile)
Donatella Castelli (OpenAIRE and CNR, Italy)
Les Carr (Southampton University, UK)
Leslie Chan (University of Toronto at Scarborough, Canada)
Chuck Humphrey (Portage, Canada)
Rick Johnson (SHARE and University of Notre Dame, US)
Petr Knoth (Jisc and Open University, UK)
Paolo Manghi (CNR, Italy)
Lazarus Matizirofa (NRF, South Africa)
Pandelis Perakakis (Open Scholar, Spain)
Jochen Schirrwagen (University of Bielefeld, Germany)
Daisy Selematsela (UNISA, South Africa)
Kathleen Shearer (COAR, Canada)
Tim Smith (CERN, Switzerland)
Herbert Van de Sompel (Los Alamos National Laboratory, US)
Paul Walk (EDINA and Antleaf, UK)
David Wilcox (Duraspace and Fedora, Canada)
Kazu Yamaji (NII, Japan)

Introduction

The widespread deployment of repository systems in higher education and research institutions provides the foundation for a distributed, globally networked infrastructure for scholarly communication. However, repository platforms are still using technologies and protocols designed almost twenty years ago, before the boom of the Web and the dominance of Google, social networking, semantic web and ubiquitous mobile devices. This is, in large part, why repositories have not fully realized their potential and function mainly as passive recipients of the final versions of their users' conventionally published research outputs. In order to leverage the value of the repository network, we need to equip it with a wider array of roles and functionalities, which can be enabled through new levels of web-centric interoperability.

In April 2016, COAR launched the Next Generation Repositories Working Group to identify the core functionalities for the next generation of repositories, as well as the architectures and technologies required to implement them. This report presents the results of work by this group over the last 1.5 year. The report describes 11 behaviours for the next generation of repositories, as well as the recommended technologies, standards and protocols that repository platforms need to incorporate in order to support these behaviours.

The work of the Next Generation Repositories Working Group was guided by a vision, principles and design assumptions included here.



Wassily Kandinsky, The Great Gate of Kiev, 1928

Vision

"Our 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."

Guiding Principles

Distribution of control

Distributed control, or governance, 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. Distributed networks are more sustainable and at less risk to buy-out or failure.

Inclusiveness and diversity

Different institutions and regions have unique and particular needs and contexts (e.g diverse language, policies and priorities). A distributed network of repositories will aim to reflect and be responsive to the different needs and contexts of different regions, disciplines and countries.

Public good

The technologies, architectures and protocols adopted in the context of the global network for repositories will be available to everyone, using global standards when that are available.

Intelligent openness and accessibility

Scholarly resources, will be made openly available and in accessible formats, whenever possible, in order increase their value and maximize their re-use for the benefit for scholarship and society.

Sustainability

Institutions and research organizations will be major participants in the global network, contributing to the long term sustainability of resources.

Interoperability

Repositories will adopt common behaviours, functionalities and standards ensuring interoperability across institutions and enabling them to engage in a common way with external service providers

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

Evolution, not revolution

We prefer to evolve solutions, adjusting existing software and systems that are already widely deployed across the world to better exploit the ubiquitous Web environment within which they are situated.

Convention over configuration

We favour the adoption of widely recognised conventions and standards, and encourage everyone to use these where possible, rather than accommodating richer, more complex and varied approaches. New standards should be introduced only when concrete and pragmatic needs arise, with the intention of 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, we want to integrate tools into the environments and systems where users are already engaged.

Characteristics of Next Generation Repositories

The next generation repository provides access to a wide diversity of resources, including published articles, pre-prints, datasets, working papers, images, software, and so on.

Resource centric

The next generation repository is resource-centric, making resources the focus of its services and infrastructure. In a global network of next generation repositories, distributed and diverse resources are openly accessible and unambiguously identified by HTTP(S) URIs[3] rather than exposed through imprecise descriptive metadata. Resources are discretely exposed, portable, networked, and pluggable in a common way, presenting a rich content layer that serves as the foundation for the development of value added services, like peer-review, social networking, recommender systems, usage measures, and so on. By becoming resource-centric in this way, repositories are established as important systems managing nodes in the global network of scholarly resources.

Networked

The next generation repository is a networked repository. Cross-repository connections are established by introducing bi-directional links as a result of an interaction between resources in different repositories, or by overlay services that consume activity metadata exposed by repositories. Links between resources in distributed repositories will create a scholarly web within the larger web and will be a key catalyst towards effectively bridging scholarly communication and research infrastructures, removing the separation between the places where we perform science and the places where we publish it. This brings many new opportunities for broadening the scope of the services repositories offer.

Machine-friendly

The next generation repository is machine-friendly, enabling the development of a wider range of global repository services, with less development effort. As opposed to current repositories, where metadata of scholarly outputs are machine accessible only through batch harvesting, the next generation repository supports machine access to the full variety of its resources using batch, navigation and notification access mechanisms.

Active

The next generation repository is active and supports versioning, commenting, updating and linking across resources. The content in the repository is not static, but will change

[3] HTTP URIs, in the web architecture, have been used to denote documents -- "web pages" informally, or "information resources" more formally. However, with the growth of the Semantic Web, which uses URIs to denote anything at all, the urge to use and practice of using HTTP URIs for arbitrary things grew steadily.

over time. The next generation repository will not just passively wait to be harvested, but they will actively notify interested systems of changes in their resources and their usage.

Conceptual Model

In our conceptual model, we draw a distinction in four areas: “content”, “overlay content”, “descriptive metadata” and “activity metadata”. We envision a system in which the relationships between a resource (content), and review or comments about that resource (overlay content), are linked through a common vocabulary and url that expresses the relationship between these two resources. Many of the behaviors and recommendations for next generation repositories pertain to establishing links across repositories as a way to break down the silos and arrive at an environment characterized by interconnected networked repositories.

The image below illustrates these relationships.

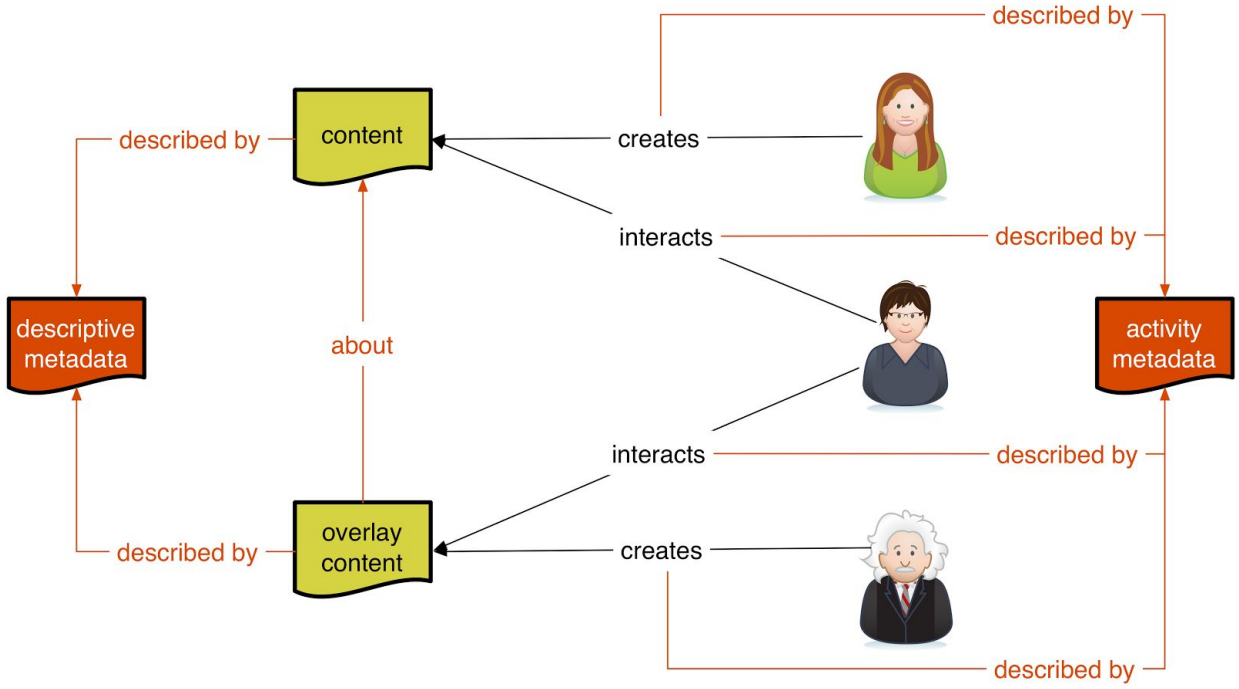


Image by Herbert Van de Sompel

Behaviours of Next Generation Repositories

In February 2017, the Working Group published several user stories that outlined the group's priority functionalities for repositories in the future, for public review and comment. The user stories were updated based on input from the community, and, in turn, were used to identify 11 new behaviours for next generation repositories. This report describes each behaviour and lists the technologies, protocols and standards recommended by the Next Generation Repositories Working Group for adoption to support each behaviour.

The new behaviours and technologies proposed here will facilitate the development of new services on top of the collective network, including social networking, peer review, notifications, and usage assessment.

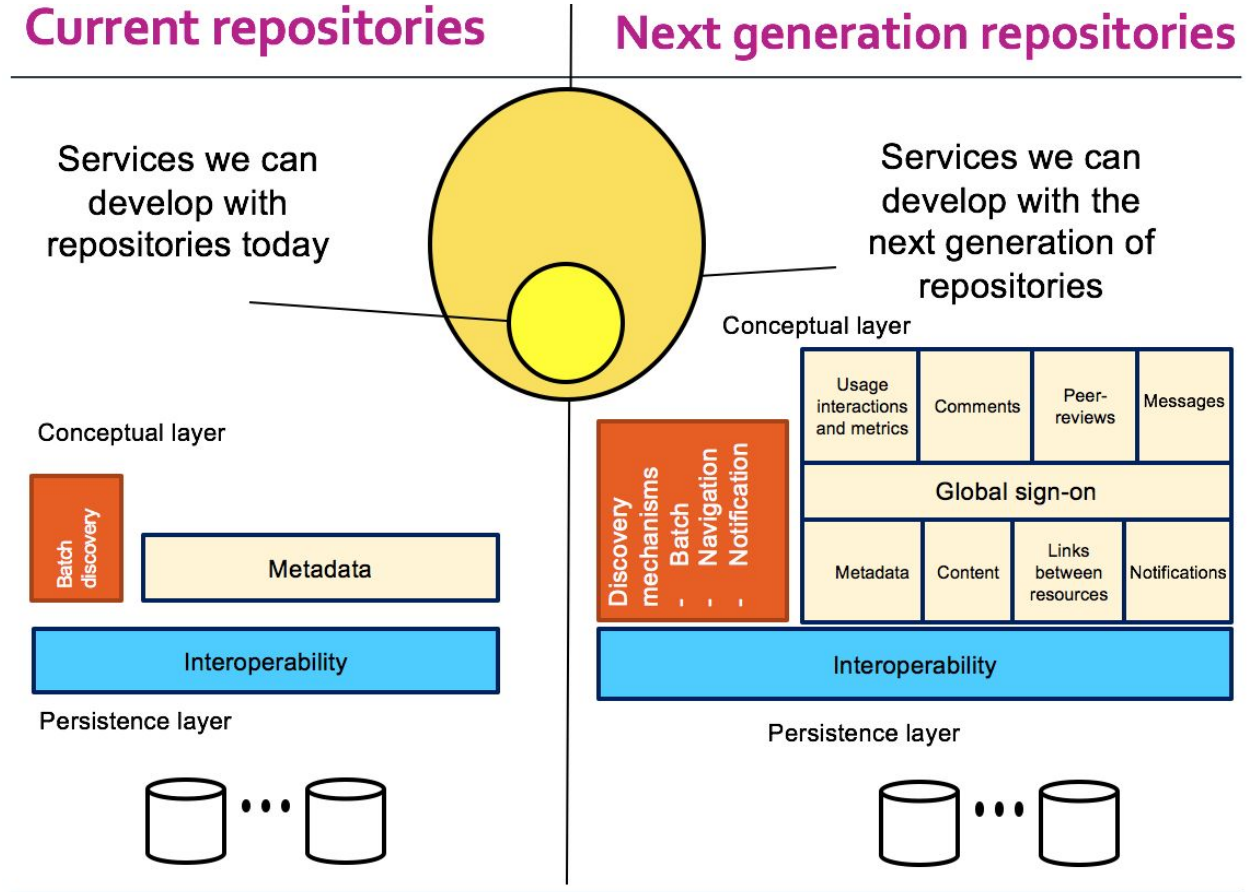


Image by Petr Knoth

The Next Generation Repositories Working Group has explicitly focused on the generic technologies required by all repositories to support the adoption of common behaviours. However, we also recognize that there are other technologies and standards that may be useful for certain content types or disciplinary communities.

In some cases, the technologies required to support a specific behaviour are not yet sufficiently mature, or it is not yet clear what technology will prevail. In other cases, where there are not currently no appropriate technologies to support the specific behaviour. In these cases, the Working Group was not able to recommend specific technologies, however we will continue to monitor developments and make recommendations as new or better technologies become available.

The behaviours and technologies are a snapshot of the current status of technology, standards and protocols available. However, we are aware that technologies will continue to evolve. To that end, we will soon be publishing the behaviours and technologies in a GitHub repository to support updates, as well as enabling greater input and engagement with the broader community as technologies evolve or new technologies come onto the scene.

List of Behaviours

1. Exposing Identifiers
2. Declaring Licenses at a Resource Level
3. Discovery through Navigation
4. Interacting with Resources (Annotation, Commentary and Review)
5. Resource Transfer
6. Batch Discovery
7. Collecting and Exposing Activities
8. Identification of Users
9. Authentication of Users
10. Exposing Standardized Usage Metrics
11. Preserving Resources

Behaviours and Recommended Technologies, Standards and Protocols

1. Exposing Identifiers

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(S) URI. The persistent HTTP(S) 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 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(S) 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(S) 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 (cite-as) to point from web resources that are part of a scholarly object to their persistent HTTP(S) 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.

User stories related to the behaviour

As a web reference manager, annotation tool, or crawler, when I encounter a landing page or any other web resource that is part of a scholarly object, I need to easily identify the associated persistent HTTP URI for the resource, so that I can retrieve it.

Technologies, standards, and protocols supporting this behaviour

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 <link> 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

2. Declaring Licenses at the Resource Level

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. 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 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.

User stories related to this behaviour

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.

Technologies, standards and protocols supporting this behaviour

Creative Commons Copyright Licenses give everyone from individual creators to large companies and institutions a simple, standardized way to grant copyright permissions to their creative work. The combination of our tools and our users is a vast and growing digital commons, a pool of content that can be copied, distributed, edited, remixed, and built upon, all within the boundaries of copyright law. https://creativecommons.org/licenses/

Signposting [see behaviour #1. Exposing Identifiers]

3. Discovery through Navigation

A scholarly object presents itself on the web as a bundle of resources, each with its own HTTP(S) URI. For example, there is the landing page, the PDF and/or HTML version of a paper, one or more supporting dataset, a bibliographic description of the scholarly object in one or more formats, etc. While a human user can intelligently move around between these various resources, understanding that they pertain to the same scholarly object, a machine can not. For example, most repositories provide links to bibliographic information that describes a scholarly object using links in the landing page discriminated by tags that identify a citation format such as “bibtex”, “RIS”, “DC”, etc. 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. These tools need to resort to repository-specific heuristics when trying to accomplish this task. Also, when these tools land on resources other than the landing page – say the PDF or the dataset - they cannot navigate to other resources that pertain to the scholarly object. In order to improve the discoverability of resources through navigation in repositories, the fact that a scholarly object is a bundle of web resources needs to be conveyed to machine agents. This can be achieved by using typed HTTP links with appropriate link relation types and format indicators to interlink the web resources that make up a scholarly object.

User stories related to this behaviour

As a human or machine user, I want to easily and uniformly discovery the metadata in a repository record, so that I can ascertain the relevance of the resource.

As a repository manager, I want to be able to access the metadata in my repository in real time through an API in order to build views or services on any platform using the data.

Technologies, standards and protocols supporting this behaviour

Signposting [see behaviour #1. Exposing Identifiers]

4. Interacting with Resources (Annotation, Commentary, and Review)

Repositories can increase their value by supporting commentary, annotation and peer review activities. The functionality to allow these activities does not necessarily need to be provided by the repositories themselves but can rather be provided by third party services or tools that specialize in the creation of overlay content. By supporting the creation of overlay content in this manner, repositories can begin to reposition themselves to the centre of scholarly communication and promote discussion and collaborative work. Achieving a level of interoperability between repositories and such third party services is essential, especially with regard to the manner in which overlay content is expressed, and the way in which the repository is made aware that overlay content was created. This allows the repository to surface the overlay content by linking to it, by ingesting it, and by exposing it to aggregators. In order to be able to unambiguously connect overlay content with its creator, global identification and authentication of users that generate it is essential (see “Identification of Users” and “Authentication of Users”).

User stories related to the behaviour

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 other readers, so that the quality of a resource can be assessed by others.

As a researcher, I want to connect content from different repositories to create meaningful aggregation such as study paths or virtual reconstruction combining separated and distributed digital objects (images, 3d objects).

As a funding institution, I want to be able to access the reviews (and metrics) of resources created by specific authors.

Technologies, standards and protocols supporting this behaviour

Activity Streams 2.0 is an approach to describe interactions with resources, including commenting, liking, sharing, etc. Interactions are expressed as JSON-LD and use the Activity Streams 2.0 vocabulary. While this core vocabulary is targeted at general social web activities, extensions can be created to supported scholarly use cases.
https://www.w3.org/TR/activitystreams-core/ ;
https://www.w3.org/TR/activitystreams-vocabulary/

Web Annotation Model and Web Annotation Protocol specify an approach to express annotations (including commentary, review, etc.) and an associated protocol to create and manage them. Annotations are expressed using an RDF-based vocabulary and can be rendered as JSON-LD. The protocol is based on HTTP and adheres to REST design

principles. https://www.w3.org/TR/annotation-model/ ; https://www.w3.org/TR/annotation-protocol/

International Image Interoperability Framework (IIIF) is a family of APIs that enable easy reuse, share and interaction with images for annotation, transcription, composing, authenticated access, etc. Despite to be a technology relevant for specific kind of content in the repository we believe it is a good example of technology to highlight to emphasize the distributed nature of the Next Generation Repositories. http://iiif.io/

With regard to technologies aimed at informing a repository that overlay content was created, and the manner in which a repository can expose this information, see behaviour #7. Collecting and Exposing Activity Metadata.

5. Resource Transfer

The vision for next-generation repositories strongly emphasises a resource-centric paradigm, where resources are not arbitrarily copied from system to system but are, rather, referenced where they are. However, there are use cases where the copying of resources (metadata, content or both) is necessary, generally to avoid the problem of network latency, to support functions which operate simultaneously on large numbers of resources, where those resources are distributed across many repositories.

Repositories should consider supporting by value content transfer of their resources to support text/data mining and preservation applications. By value content transfer entails allowing third parties to efficiently access and transfer the actual content of scholarly objects. When text/data mining and preservation activities are carried out in infrastructures external to the repository, the custodians of these infrastructures need to be able to transfer the content over from the repository in an efficient and timely manner. This includes being able to recurrently synchronise their holdings with that of the repository as its resources evolve (created/updated/deleted). This can be achieved in a by reference manner by exposing a list of URIs of resources in the repository (see "Batch Discovery"), but that approach can become problematic for larger repositories. A by value approach for content transfer in which both content as metadata are exposed is more appropriate in such cases.

User stories related to the behaviour

As a human or machine user, I want to be able to mine the collective full text content of repositories to discover new relationships and make new discoveries.

Technologies, standards and protocols supporting this behaviour

IPFS is a promising emerging peer-to-peer hypermedia protocol aimed at making the web faster, safer, and more open. IPFS should be considered as a possible approach for cases where large data collections need to be shared among a number of parties, each of which actively operates an IPFS node. https://ipfs.io/

ResourceSync is a specification based on Sitemaps that can be used by repository managers to provide information that allows third-party systems to remain in sync with the resources in their repository as they evolve, i.e. are created, updated, deleted. Whereas basic Sitemaps allow exposing a repository inventory and crawl-related metadata, ResourceSync adds ways to expose changes only, and to provide expressive synchronization-related metadata as well as typed links for further discovery. ResourceSync can be used for discovery and synchronization of both content and metadata and uses the Sitemaps XML format.

SWORD (Simple Web-service Offering Repository Deposit) is a lightweight protocol for depositing content from one location to another. It stands for Simple Web-service Offering Repository Deposit and is a profile of the Atom Publishing Protocol.
http://swordapp.org/about/

6. Batch Discovery

Uniform, global, cross-repository discovery of resources is essential to establish repositories as important players in scholarly communication. Batch discovery generally supports search, but also use cases that require content transfer such as text mining and preservation. The better resources in repositories are surfaced using batch discovery mechanisms, the more likely they are to be found by users and applications alike. Supporting batch discovery to enable specialized services avoids the problem of "if it did not appear near the top of a results list, it does not exist."

User stories related to this behaviour

As a user, I want to discover repository materials of interest via aggregators or other search services such as BASE, CORE, OpenAIRE, and so on.

A text mining application wants to discover the HTML or PDF versions of scholarly publications.

A digital preservation application wants to discover all resources that pertain to a scholarly object, including all its constituent resources in various representations, bibliographic information, license information, and a persistent identifier.

Technologies, standards and protocols supporting this behaviour

ResourceSync [see behaviour #5. Resource Transfer]

Signposting [see behaviour #1. Exposing Identifiers]

Sitemaps are widely used by webmasters to inform search engines about pages on their sites that are available for crawling. In its simplest form, a Sitemap is an XML file that lists a URL for each available resource along with optional additional metadata about that resource aimed at optimizing the crawling process (e.g. last modified date, estimated change frequency). Repository managers can use Sitemaps as a straightforward way to expose a repository inventory to search engines. https://www.sitemaps.org/

7. Collecting and Exposing Activities

Repositories should be able to actively and in real-time collect and expose activity (e.g. information about changes, additions, comments, annotations, peer-reviews, accessess, downloads, etc.) pertaining to scholarly objects they host. Authors of the scholarly object involved in an activity, other repositories, and a variety of consuming applications that keep the pulse on scholarship as it happens should be able to receive metadata about activity not only retrospectively through harvesting, but also in real-time. To that end, notification mechanisms need to be put in place. Depending on the use case, these could be point-to-point notifications (e.g. an author is directly notified about a citation to her paper) or publish/subscribe notifications (e.g. a consuming application interested in peer-review subscribes to a channel on which review events are posted). In addition, value added services should be able to consume such activity information producing new notifications in turn. For example, this could be exemplified by academic recommender systems, which can, based on past (even anonymous) activity information, significantly help users in navigating research objects stored across repositories globally. In order to achieve such functionality, unique identification (by means of HTTP(S) URIs) of scholarly objects and actors (e.g. authors, reviewers, institutions) in the scholarly communication environment is essential.

User stories related to the behaviour

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.

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.

As a repository manager I want other systems to be notified of changes made to my collection to ensure that records are standardized across various locations.

As an author, I want to be informed as soon as my paper gets cited, my dataset is re-used, etc.

A a repository manager, I want to know when web resource link to resources in my repository. That way, I can create links back to those resources and support discovery of related resources.

Technologies, standards and protocols supporting this behaviour

Activity Streams 2.0 [See behaviour #4. Interact with Resources]

Linked Data Notifications is a general purpose notification protocol whereby any resource can advertise an inbox to which notifications pertaining to that resource can be posted. For example, an annotation, commenting, or reviewing application can post a notification to a resource's inbox to inform that resource that an interaction occurred with

it, what the nature of the interaction was, who the actor involved in the interaction was, etc. The payload of a notification is expressed as JSON-LD and uses the Activity Streams 2.0 vocabulary. A repository could support an inbox per resource, or an inbox for the entire repository. The repository could surface interactions that took place with its resources in the user interface, could further post them to the inbox of an aggregating application, or could expose them in the aggregate for further machine consumption using WebSub (see below). https://www.w3.org/TR/ldn/

ResourceSync Change Notifications is a publish/subscribe protocol based on WebSub and focused on sending notifications about changes (create/update/delete) to resources in a repository to subscribers. ResourceSync Change Notifications can be used for discovery and synchronization of both content and metadata and use the Sitemaps XML format. http://www.openarchives.org/rs/notification

Signposting [see behaviour #1. Exposing Identifiers]

Webmention is a simple, point-to-point, trackback/pingback approach aimed at informing a resource that it was linked from another resource. It allows, for example, the establishment of bidirectional links. https://www.w3.org/TR/webmention/

WebSub is a publish/subscribe protocol, whereby a publisher posts resource updates to a channel on a hub and the hub subsequently relays those updates to channel subscribers. A repository could publish interactions that took place with its resources on a single channel, or on multiple channels, for example, one per type of activity (e.g. citation, review, annotating). This could be achieved in a manner similar to what is specified for ResourceSync Change Notifications. Aggregating applications could (selectively) subscribe to these repository channels. https://www.w3.org/TR/websub/

Other messaging protocols (e.g. AMQP, Kafka) provide a common mechanism for communication between publishers of any kind of Web content and their subscribers

We also need to expose user interaction data in standard format and with a common vocabulary.

8. Identification of Users

Repositories should support the creation of overlay content such as annotation, commentary, peer review, as well as other interactions with the scholarly objects they host. Inviting users to identify themselves by means of identifiers that have global reach (HTTP(S) URIs) when interacting with objects in this manner can lead to constructive conversations and the creation or reinforcement of social connections. User identification can support personalized services such as targeted notifications and recommendation systems that help users to more efficiently navigate large-scale distributed collections. Overall, we need the ability to uniformly identify users, i.e. the ability to understand that particular activities performed in any of the repositories in the network belong to the same user (regardless of whether the user is authenticated or not). This will add a global dimension to repositories and help to move beyond the status quo that is perceived to be largely silo-ed. We also want to record activities of anonymous users to better understand how content across the global repositories network is consumed.

User stories related to this behaviour

As a user, I want my repository to recognize me and others 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.

As a 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.

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.

As a 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.

As a user, I want to know when 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.

As a user, 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.

Technologies, standards and protocols supporting this behaviour

ORCID is an HTTP(S) URI in the orcid.org domain aimed at unambiguously identifying a scholarly contributor. ORCIDs are increasingly used in a variety of scholarly workflows. A profile is associated with a contributor's ORCID, which has both a human and machine

readable representation. The machine-readable profile is RDF-based and uses the FOAF vocabulary. The ORCID organization also provides authentication services that can be used in distributed settings, see “Authentication of Users”. https://orcid.org/

Social Network Identities are provided by several social network platforms. In many cases, these platforms also provide facilities for distributed authentication based on the social network identities they provide as described in behaviour “#9. Authentication of Users”.

WebID is an HTTP(S) URI which refers to an agent (person, organization, group, etc.) and that is minted in a domain that is typically owned by the agent. The WebID leads to a machine-readable profile that describes the agent. The RDF-based profile is fully under the agent’s control and uses the FOAF vocabulary. A WebID is commonly used in conjunction with the WebID/TLS authentication approach (see behaviour “#9. Authenticating Users”) and the Web Access Control Lists authorization approach.
https://www.w3.org/2005/Incubator/webid/spec/identity/

9. Authentication of Users

Requiring users to identify themselves by means of identifiers that have global reach (HTTP(S) URIs) when interacting (e.g. annotation, commentary, review) with scholarly objects hosted by a repository can lead to constructive conversations and the creation or reinforcement of social connections. Overall, the ability to uniformly identify users that interact with content hosted in repositories, worldwide, will add a global dimension to repositories and help to move beyond the status quo that is perceived to be largely silo-ed. But providing a global identity when interacting with repository content is not sufficient. The identity that a user claims must be verified with the provider that assigned the identity to the user in the first place. Therefore, repositories must support approaches that allow verification of identities provided by users, both for academic identities (i.e. ORCID) and identities provided by social networks (e.g. Twitter, Google, Facebook, Mastadon).

User stories related to this behaviour

As a user, I want the repository to recognize me and others 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.

As a repository manager, I want to avoid that users interact in inappropriate ways with content in my repository. Requiring users to identify themselves and verifying the claimed identity with the identity provider reduces the risk.

Technologies, standards and protocols supporting this behaviour

HTTP Signatures provide an authentication approach that is conceptually similar to WebID/TLS. But the approach is more generic in that it is not solely tied to the WebID concept. Also, in addition to authentication, it allows verification that the communication between client and server was not tampered with. The approach is currently being standardized at the IETF and is definitely something to keep an eye on in the authentication space. https://datatracker.ietf.org/doc/draft-cavage-http-signatures/

OpenID Connect 1.0 is a simple identity layer on top of the OAuth 2.0 protocol, which itself is used for distributed authentication against compliant identity providers. OpenID Connect allows client applications - such as repositories and browsers - to verify a user's claimed identity by authenticating the user against her identity provider. As a result of a successful authentication, basic profile information about the user can be passed along to the client application. The specification is extensible, allowing participants to use optional features such as encryption of identity data, discovery of OpenID Providers, and session management. The major providers of social network identities already support OpenID Connect. ORCID's implementation is currently in beta. http://openid.net/connect/

WebID/TLS is a protocol that enables secure user authentication on the basis of the Transport Security Layer protocol (TSL), X.509 Certificates, and a WebID with associated

profile. It enables a user to authenticate by simply choosing an appropriate certificate from the ones proposed by the browser. The certificate is used to sign a server's challenge with the user's private key but also to convey the user's WebID. The WebID leads the server to the user's profile, which contains her private key, allowing the server to verify that the challenge was met correctly. While this authentication approach is both elegant, efficient, and fully distributed, its adoption has thus far been hindered among others due to issues with generating certificates and user interface challenges.
https://www.w3.org/2005/Incubator/webid/spec/tls/

10. Exposing Standardized Usage Metrics

Repositories should be able to share user interaction data to enable the development, deployment and evaluation of innovative value-added global services over repositories. Collecting standard metrics is important in order to optimise, operate, and enhance the repository and demonstrate the value of the repository to authors and other stakeholders. Methodologies for measuring usage must be standardized across repositories and repository platforms. Measures also need to be reliable and trusted by the community as accurate so they can be compared across platforms. Additionally, when repositories host copies of the same article, they should be able share and sum their separate usage metrics, which in turn will let the author (and other users) see the overall, aggregate statistics. Perhaps most importantly, if we can create a trusted system of standardized usage metrics across the global network, we can create an alternative, journal-independent reputation system, taking away some of the influence and power of the current commercial publishers. That being said, given the inherent limitations of quantitative measures in general for assessing quality and relevance of research, the qualitative functionality of the global network as supported through annotations, reviews and comments is critical.

Exposing usage metrics can be done in either of two modes: pull mode (for example using SUSHI) or push mode by a tracking protocol to a service provider, which currently is vendor specific (for example, google-analytics, IRUS-UK, OpenAIRE using Piwik, RAMP). However, one of the main challenges for exposing usage metrics is ensuring the metrics are open and comparable, something that cannot be solved by technology alone, but rather the adoption of common standards.

User stories related to the behaviour

As an author, I want to know how often my paper, dataset or other resource is being used, 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.

As a funder, I want to use repository metrics as one measure that will help evaluate the impact of the research I fund.

As a research administrator, I wish to use a broader variety of measures to assess impact including repository metrics and incorporate them in my reports that assess the impact of the research I support.

Technologies, standards and protocols supporting this behaviour

With technologies needed to support transfer of resources into preservation platforms, see behaviour #5. Resource Transfer

COUNTER provides the standard that enables the knowledge community to count the use of electronic resources. Known as the Code of Practice, the standard ensures vendors and publishers can provide their library customers with consistent, credible and comparable usage data. https://www.projectcounter.org/

SUSHI (Standardized Usage Statistics Harvesting Initiative) is an ANSI/NISO Standard that defines automated request and response model for harvesting e-resource usage data. It is designed to work with COUNTER, the most frequently retrieved usage reports.

ETag or entity tag is part of HTTP, the protocol for the World Wide Web. It is one of several mechanisms that HTTP provides for web cache validation, which allows a client to make conditional requests. This allows caches to be more efficient, and saves bandwidth, as a web server does not need to send a full response if the content has not changed. ETags can also be used for optimistic concurrency control, as a way to help prevent simultaneous updates of a resource from overwriting each other. This is relevant to support central systems from fetching only new data about metrics.
https://en.wikipedia.org/wiki/HTTP_ETag

Usage metrics service provider for repositories (IRUS-UK http://irus.mimas.ac.uk/; OpenAIRE using Piwik https://piwik.org/; RAMP - Repository Analytics and Metrics Portal http://ramp.montana.edu/

11. Preserving Resources

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. Repositories should try obtaining the most reusable format (e.g. latex, TEI rather than a PDF) by validating how manuscripts were created, such as it is currently done by arXiv.org (DDI instead of SPSS or XLSX) and encouraging the deposition of that format.

User stories related to the behaviour

As a scholar or institution, 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.

Technologies, standards and protocols supporting this behaviour

Digital preservation is the active management of digital content over time to ensure ongoing access to resources. Preservation is an extremely complex activity, involving the adoption of appropriate policies, standards, practices, and technologies. There are already dedicated communities focused on defining best practices and technologies for digital preservation, therefore the Working Group did not addressed the specific technologies in this behaviour in a comprehensive way, but rather focussed on the ability of repositories to transfer full text content from repositories to preservation platforms. The technologies to support this behaviour are already described in behaviour #4. Resource Transfer.

Next Steps

One mission of a repository is to manage and provide access to the valuable and diverse intellectual output of the community it serves. However, equally important is that repositories are nodes in a larger network, contributing their collective contents to a global knowledge commons on top of which value added services can be built.

The distributed network of repositories can and should be a powerful tool to promote the transformation of the scholarly communication ecosystem, making it more research-centric, innovative, while also managed by the scholarly community. However, this vision rests on the notion that repositories behave (or function) in common ways, and interact with external services in the same manner. As such, it is important that the technologies, standards and protocols defined here are widely accepted and adopted by repositories around the world.

COAR is committed to disseminating the technological recommendations contained in this report widely. In the coming months, COAR will work through a variety of mechanisms with different stakeholder communities (repository platform providers; libraries and institutions that maintain repositories; repository networks; and other value added service providers) to promote the adoption of new technologies as widely as possible.

Evolution of Technologies

COAR and the Next Generation Repositories Working Group are keenly aware that technologies evolve rapidly and there is a need to continually monitor technologies and developments that will support priority behaviours. In the coming weeks, we will be placing the behaviours and recommendations into a GitHub repository. This will allow the community to provide comments on the Working Group's recommendations as well as suggestions for other technologies, standards and protocols that should be considered.