NEXT GENERATION REPOSITORIES

COAR Confederation of Open Access Repositories

Towards Next Generation Repositories

Kazu YAMAJI

With significant input from Andrea Bollini, Petr Knoth, Eloy Rodrigues, Kathleen Shearer, Herbert Van de Sompel, Paul Walk, David Wilcox, and the COAR NGR Working Group

NEXT GENERATION REPOSITORIES

COAR's Vision

COAR Confederation of Open Access Repositories

A global knowledge commons based on a network of open access repositories

Who is COAR?

- An international association founded in 2009
- Members & Partners: over 120 institutions from 35 countries in Africa, Asia, Australasia, Europe, North and South America

Objectives:

- Strategic voice for repositories
- Interoperability and alignment across repositories and regional networks
- Capacity building
- Support the development of value added services

What is about Repository

- Clifford Lynch (2003)
  - a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members
- Jisc (2016)
  - A repository is a set of services[1] that a research organisation[2] offers[3] to the members of its community[4] for the management and dissemination[5] of digital materials[6] created by its community members (Provide detailed definition from 1 to 6)
- Herbert Van de Sompel (2016)
  - The purpose/mission of repositories is no longer well defined and has IMO to quite an extent drifted since their original inception from being about "**all kinds of digital materials created by an institution's staff**" to "**formally published materials created by an institution's staff**". In this drift lies (IMO) one of the major problems of many current IRs: they don't provide a service to their local community…

Your Thought ?

- Self-archive Green OA
- Gold OA with Repository

Why Now ?

- Common Standard for Repository Network
  - ⇨ only OAI-PMH
  - – Background
    - Create New Scholarly Communication
      - – by Pre-Print: arXiv
    - Based on Web-Technology around 90’s
  - – Behavior
    - Metadata Aggregation by Machine
    - Pull Action from Service Provider (Aggregator)

Next Generation Repositories

Major strategic priority for COAR
Working Group launched in April 2016

The problem: Repositories have not fully realized their potential and function mainly as passive, silo'ed recipients of the final versions of their users' conventionally published research outputs

Aim: to identify functionalities and architectures for the next generation repositories within the context of scholarly communication

Next Generation Repositories Working Group, 2016-2017

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 (Open University, Jisc, 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)

Next Generation Repositories

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.

Characteristics of NGR

- It manages and provides access to a **wide diversity of resources**, including published articles, pre-prints, datasets, working papers, images, software, and so on.
- It is **resource-centric**, making resources the focus of its services and infrastructure
- It 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
- It is **machine-friendly**, enabling the development of a wider range of global repository services, with less development effort
- It is **active and supports versioning, commenting, updating and linking across resources**

Principles of NGR, Part 1

- ***Distribution of control*** – Distributed control, or governance, of scholarly resources (preprints, 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.

Principles of NGR, Part 2

- ***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 of NGR Part 1

- ***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.

Design Assumptions of NGR Part 2

- ***Convention over configuration*** – Our preference is to **adopt widely recognised conventions and standards**, and encouraging everyone to use these where possible, rather than accommodating richer, more complex and varied approaches. As a corollary to this, we believe 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, integrate tools into the environments and systems where they are already engaged.

Methodology

1. Identify major use cases
2. Determine functionalities/behaviours
3. Develop conceptual models
4. Define technologies and architectures
5. Publish recommendations (November 2017)
6. **Support adoption and implementation**

Final Report

- Recommendations of the COAR Next Generation Repositories Working Group (November 2017)
  - https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf (pdf)
  - http://ngr.coar-repositories.org/ (website)
  - https://github.com/coar-repositories/ngr/tree/master/webroot/content/behaviour (GitHub)

Behaviors and Technical Recommendations

1. Exposing Identifiers
2. Declaring Licenses at the 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

Technologies, Standards and Protocols

1. Activity Streams 2.0
2. COUNTER
3. Creative Commons Licenses
4. ETag
5. HTTP Signatures
6. IPFS International Image
7. Interoperability Framework
8. Linked Data Notifications
9. ORCID and other author IDs
10. OpenID Connect
11. ResourceSync
12. SUSHI
13. SWORD
14. Signposting
15. Sitemaps
16. Social Network Identities
17. Web Annotation Model and Protocol
18. WebID and WebID/TLS
19. WebSub
20. Webmention

6. Batch Discovery

- User stories
  - 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 version s 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
  - ResourceSync
  - Signposting
  - Sitemaps

Next Generation Repositories

Behaviours and Technical Recommendations

Technologies, Standards and Protocols

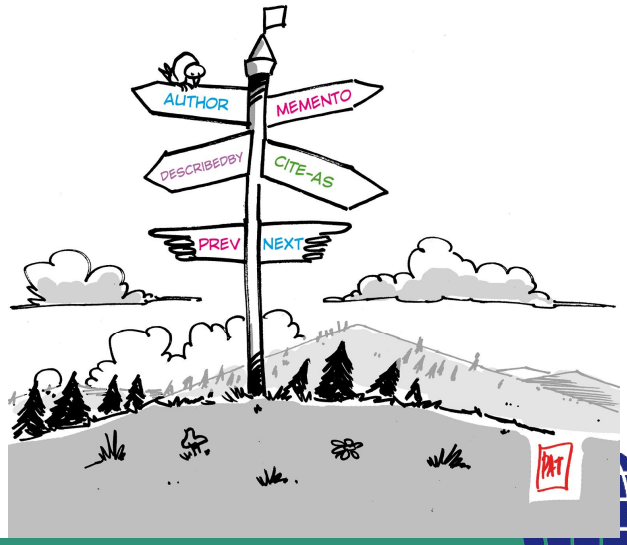
Two examples…

Signposting

An approach to make the scholarly web more friendly to machines exposing relations as Typed Links in HTTP Link headers

The following discovering patterns are currently defined:

- Author
- Bibliographic Metadata
- Identifier
- Publication Boundary



ResourceSync

- Successor of the OAI-PMH protocol and much more…
- Faster, reliable and scalable
- Allows real-time notification
  (and recovering of missed messages)
- Drives resource synchronization:
  content and metadata are both managed

http://www.openarchives.org/rs/1.1/resourcesync

Next Steps

1. Implementation of technologies in repository platforms
2. Development of network or hub services
3. Ongoing monitoring of new technologies, standards and protocols

Next Steps 1

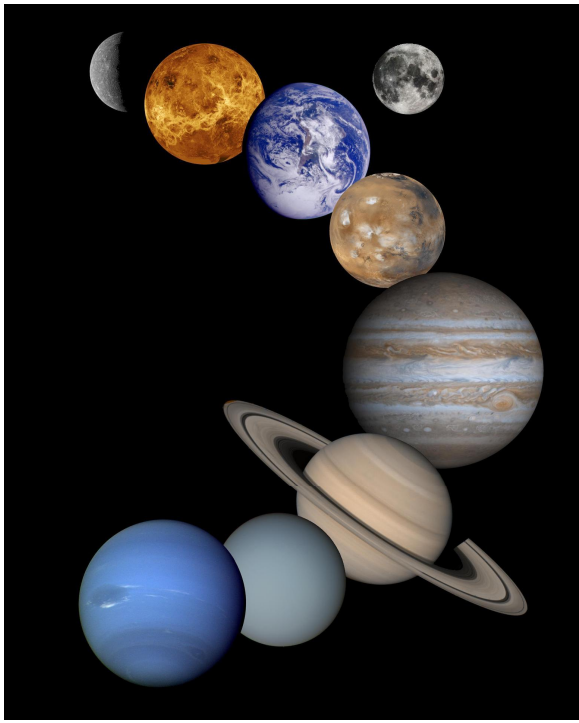
Implementation of technologies in repository platforms

- Working with open source platforms to implement recommendations
- Organizing regional projects to support development costs and contribute resources: OpenAIRE Advance (Europe), National Institute of Informatics (Japan), US NGR Implementers Group

Next Steps 2

Development of network or hub services

- Aligning Repository Networks Accord, May 2018
- Technical and Strategic Meeting of 20 Repository Networks, May 14 & 15, 2018 in Hamburg, Germany
- Pilot Projects 2nd half 2018 (Open Peer Review, Common Standards for Usage Statistics, Recommender Systems)



Next Steps 3

Ongoing monitoring of new technologies, standards and protocols

COAR Next Generation Repositories Editorial Group

-Andrea Bollini
-Rick Johnson
-Paolo Manghi
-Eloy Rodrigues
-Kathleen Shearer
-Herbert Van de Sompel
-Paul Walk
-Kazu Yamaji

Domestication of NGR

What we are doing in Japan

1. Review Japanese IR Activity in the Past
2. Summarize Next Challenge
3. Understand NGR Concept
4. Mapping Our Challenge with NGR
5. Define 12th and more Behaviors if Necessary
6. Create the Future Repository Network

NEXT GENERATION REPOSITORIES

COAR Confederation of Open Access Repositories

Thanks and questions!