



*Promoting greater visibility and application of research through global networks of  
Open Access repositories*

---

## COAR Roadmap

### Future Directions for Repository Interoperability

---

Working Group 2: Repository Interoperability

February 2015

<http://coar-repositories.org>



This work is licensed under the [Creative Commons Attribution 4.0](#) License.

## Table of Contents

Acknowledgements and Contributors.....	3
Executive Summary.....	4
I Introduction.....	6
I.1 Repositories – the historical context .....	6
I.2 Trends in scholarly communication .....	6
I.3 Strategic challenges for interoperability.....	7
2 The Preparation of the Interoperability Roadmap .....	7
2.1 Vision, goal and objectives.....	7
2.2 User requirements .....	8
2.3 Participating systems and stakeholders .....	10
3 Interoperability Issues.....	12
4 Results and Analysis .....	13
4.1 Priorities according to topic area .....	13
4.2 Priorities according to specific issues .....	14
5 Conclusion .....	19
Appendix 1: The Glossary.....	21
Appendix 2: Acronyms and Abbreviations .....	24
Appendix 3: The questionnaire and its response .....	25
Key Aspect: Impact and Visibility.....	25
Key Aspect: Data Issues .....	32
Key Aspect: Validation and Aggregation .....	41
Key Aspect: Usability .....	45
Key Aspect: Sustainability .....	56
Key Aspect: Technical Issues .....	61
Issues Scatter Diagrams .....	66
Overview Key Aspects.....	67
About COAR .....	75

## Acknowledgements and Contributors

The preparation of this Roadmap was made possible because of the contribution and expertise of many people who provided input and feedback on the topic interoperability of repositories throughout a year long process.

Starting with a discussion and comments from an Expert Advisory Panel, the editorial group derived a comprehensive list of relevant interoperability issues and prepared a questionnaire. The questionnaire was then disseminated among the panel for feedback about general relevance, time factor and level of complexity. Their feedback has been used to develop a priority list of interoperability issues.

### Lead Editors:

- **Friedrich Summann**, Bielefeld University, Germany
- **Kathleen Shearer**, Confederation of Open Access Repositories (COAR), Canada

### Editors:

- **Timo Borst**, Leibniz Information Center for Economics, Germany
- **Pablo de Castro**, EDINA National Data Centre Edinburgh, UK
- **Wolfram Horstmann**, University of Göttingen, Germany
- **Alicia López Medina**, National Distance Education University Madrid, Spain
- **Katharina Müller**, University of Göttingen, Germany
- **Maxie Putzlitz**, University of Göttingen, Germany
- **Eloy Rodrigues**, University of Minho, Portugal
- **Jochen Schirrwagen**, Bielefeld University, Germany

### Experts and Reviewers:

- **Isidro Aguillo**, CINDOC-CSIC, Spain
- **Ana Alice Baptista**, University of Minho, Portugal
- **Tom Beirender**, World Bank Group, USA
- **Daniel Beucke**, University of Göttingen, Germany
- **Sheridan Brown**, V4OA Project Consultant, UK
- **Donatella Castelli**, Italian National Research Council, Italy
- **Gernot Deinzer**, University of Regensburg, Germany
- **Patrick Hochstenbach**, Ghent University, Belgium
- **Maarten Hoogerwerf**, Data Archiving and Networked Services (DANS), The Netherlands
- **Keith G. Jeffery** Consultant, UK
- **Johannes Keizer**, Food and Agriculture Organization of the United Nations, Italy
- **Thomas Krichel**, Long Island University, USA
- **Clifford Lynch**, Coalition for Networked Information (CNI), USA
- **Devika Madalli**, Indian Statistical Institute, India
- **Salvatore Mele**, CERN, Switzerland
- **Susan Reilly**, LIBER (Association of European Research Libraries), The Netherlands
- **Frank Scholze**, Karlsruhe Institute of Technology (KIT), Germany
- **Miguel Ángel Sicilia**, University of Alcalá, Spain
- **Paul Vierkant**, Humboldt University of Berlin, Germany
- **Paul Walk**, University of Bath, UK

## Executive Summary

In the past few years, Open Access repositories and their associated services have become an important component of the global e-research infrastructure. Increasingly, repositories are also being integrated with other systems, such as research administrative systems and with research data repositories, with the aim of providing a more integrated and seamless suite of services to various communities. Repositories can also be connected into networks (e.g. at the national or regional level) to support unified access to an open, aggregated collection of scholarship and related materials that machines can mine enabling researchers to work with content in new ways and allowing funders and institutions to track research outputs.

Scholarly communication is undergoing fundamental changes, in particular with new requirements for open access to research outputs, new forms of peer-review, and alternative methods for measuring impact. In parallel, technical developments, especially in communication and interface technologies facilitate bi-directional data exchange across related applications and systems. The aim of this roadmap is to identify important trends and their associated action points in order for the repository community to determine priorities for further investments in interoperability.

The roadmap process began with the compilation of a comprehensive list of interoperability issues derived from a broad discussion in the information, publishing and repository community. An Expert Advisory Panel was then asked to rate each issue according to its level of complexity and temporal relevance (or timing). This report presents the results of this process, ranking the issues according to these dimensions. The table below presents the key aspects in a two-dimensional structure.

	<b>Short term</b>	<b>Medium term</b>	<b>Long term</b>
<b>Low Complexity</b>	<ul style="list-style-type: none"> <li>• Exposing Citation Formats</li> <li>• Supporting Data Export Functions</li> <li>• Supporting Author Identification Systems</li> <li>• Supporting Search Engine Optimization (SEO)</li> <li>• Exposing Publication Lists</li> <li>• Integrating Different Persistent Identifiers</li> </ul>	<ul style="list-style-type: none"> <li>• Exposing Persistent Identifiers</li> <li>• Supporting Authorization and Authentication</li> <li>• Improving Platform Stability</li> <li>• Supporting Institutional Services</li> <li>• Extending End-User Usability</li> <li>• Validating Repository Metadata</li> <li>• Supporting Visibility in Repository Registries</li> <li>• Supporting OAI Service Provider Usage</li> <li>• Integrating Availability Services</li> <li>• Supporting Embedding Services</li> <li>• Supporting Repository Ranking Systems</li> </ul>	
<b>Moderate Complexity</b>	<ul style="list-style-type: none"> <li>• Exposing Bibliometric Information</li> </ul>	<ul style="list-style-type: none"> <li>• Exposing Versioning Information</li> <li>• De Duplication</li> <li>• Improving Registry Infrastructure</li> </ul>	

		<ul style="list-style-type: none"> <li>• Monitoring Open Access Mandate Compliance</li> </ul>	
<b>High Complexity</b>	<ul style="list-style-type: none"> <li>• Exposing Usage Statistics</li> <li>• Supporting Additional Metadata Format(s)</li> </ul>	<ul style="list-style-type: none"> <li>• Publication of Research Data</li> <li>• Improving Metadata Quality (Data Curation)</li> <li>• Processing Related Full Text</li> <li>• Supporting Deposit Protocols</li> <li>• Defining Architectural Recommendations for Repositories and their Interoperability</li> <li>• Supporting Enhanced Publications</li> </ul>	<ul style="list-style-type: none"> <li>• Extending Usage of Visualization Tools</li> <li>• Supporting Linked (Open) Data</li> <li>• Extending/Replacing Metadata Exposition Protocols</li> <li>• Handling of Complex/Compound/Nested Repository Objects</li> <li>• Supporting Long-term Preservation and Archiving</li> </ul>

Through this process, nine issues have been identified as having immediate relevance, with varying levels of complexity. These issues can be viewed as represent the most pressing priorities for efforts around interoperability.

COAR is already working to advance interoperability in several of the priority areas including author identification systems, publication lists, persistent identifiers, usage statistics and bibliometric formats. In the fall of 2014, COAR launched an international working group with the major regional repository networks, as well as CASRAI and EuroCRIS to develop a blueprint for interoperability with the aim of developing a formal mechanism whereby these interoperability issues can be discussed and addressed.

Still, many challenges remain with improving interoperability. Many of the nine issues involve some level of standardization across vocabularies, metadata and indicators, both within the repository environment as well as with other systems. Interoperability in these areas, therefore, will require collaboration across countries and regions as well as with other systems developed by different communities. In order to achieve interoperability, the repository community must work with and engage in ongoing dialogue with these other communities. In addition, ensuring local implementation of guidelines and standards at the level of individual repositories is very difficult and often requires significant community outreach to raise awareness of the benefits of adopting standards. One strategy is to work with the repository platform developers to have the standards implemented into repository software systems. In parallel the available interfaces of repositories and the corresponding systems should be open to enable bi-directional communication and information channels in order to allow concrete system interoperability.

Despite the challenges, the success of future repository services depends on the seamless alignment of the diverse stakeholders at the local, i.e. institutional, national and international community level. COAR, with its vision of a global network of open access repositories, will continue to work towards greater interoperability both within the repository community as well as with other players in the scholarly communication system.

## I Introduction

### 1.1 Repositories – the historical context

Institutional open access repositories have been around for over 12 years now, along with the corresponding definition for OAI-PMH, which was released in 2001. To date, there are more than 3000 repositories world-wide<sup>1</sup>, with a collective coverage of more than 100 million objects of varying content types and quality. During the initial years of adoption, the major focus for institutional repositories was to provide a local window to the content produced at the institution, with services developed around supporting this institutional perspective. Interoperability efforts focused on the exchange of metadata via OAI-PMH, and third-party services that were established mainly to aggregate repository metadata or expose the repositories' web pages so they can be indexed by search engines.

Over a decade later, repositories have become well-established infrastructure components and have adopted a variety of roles in the scholarly communication environment, including research assessment, open access, publishing, and preservation. In each of these roles, repositories interact with each other and other systems, requiring some level of interoperability across systems. COAR recognizes that, “The real value of repositories is their potential to be connected in order to develop a network of repositories which enables unified access to an open, aggregated mass of scholarship and related materials that machines and researchers can work with in new ways.”<sup>2</sup>

### 1.2 Trends in scholarly communication

Enabled by rapid developments in information and communication technologies, the scholarly communication system is undergoing a fundamental shift towards new methods, services and tools that support the concept of open science. Open science requires seamless access, use, reuse and trust in the validity of research outcomes (which include publications and research data, but also methodologies, software and hardware). Contextualization of research results and linking of publications, datasets and project information are fundamental aspects of this new environment, as is data centric science characterized by the large-scale processing of massive datasets. In addition, new forms of open peer review and new impact measures for research outputs are developing rapidly. The full consequences of these transformations are not yet clear, but undoubtedly they will have a strong impact on repositories and their role.

One likely future scenario for repositories is that they will become largely invisible to the user, but act as a background service supporting external service providers and visibility of content via (academic) search engines, aggregators, indexes and so on. Thus, quality of content, metadata, value added services and reliability of service will be key requirements in order to fulfill end-users needs and remain competitive with similar services. This invisibility comes with an inherent danger that repositories will be overlooked and one cannot rule out the possibility that the repository as a service will be absorbed by other infrastructure elements or will eventually be integrated into other services.

<sup>1</sup> OpenDOAR. [www.opendoar.org](http://www.opendoar.org) lists 2728, BASE [www.base-search.net](http://www.base-search.net) 3268 on Nov 25, 2014

<sup>2</sup> COAR. Current State of Repository Interoperability. 2012. pg. 4 Available at: <https://www.coar-repositories.org/files/COAR-Current-State-of-Open-Access-Repository-Interoperability-26-10-2012.pdf>

For repositories to remain relevant in this rapidly changing environment, we as a community must adopt a perspective of responsiveness, adaptability, and focus on developing services of value to the research community and other users. Short distances to the scientist and close relationships with the academic institution in combination with service orientation are examples of home-field advantages which repositories could use to hold and strengthen their position. Repositories must also maintain the balance of planning to address the future needs with supporting present practical problems and improving the actual quality level.

### 1.3 Strategic challenges for interoperability

As noted in COAR's 2012 State of Repository Interoperability Report, repository infrastructure is evolving rapidly and has lead to an interoperability landscape that seems quite "chaotic, confusing, and complex"<sup>3</sup>. In order to avoid a situation where repositories behave as local silos, a major focus for repositories must be to ensure that their content and repository systems are interoperable. Therefore increasingly relevant for the repository community will be the adoption of common metadata, identifiers (for authors, institutions, research funding organizations, publications), vocabularies and taxonomies.

In addition, repositories exist in an increasingly complex ecosystem which includes interaction with other local infrastructures and external systems dealing with publications (e.g. collaborative research environments, e-learning systems, publishing platforms, CRIS systems, etc.). This creates the need to extend interoperability activities beyond repository-to-repository efforts to include interoperability across the diversity of systems that exist in this ecosystem.

## 2 The Preparation of the Interoperability Roadmap

### 2.1 Vision, goal and objectives

COAR's vision is that "researchers, regardless of location or discipline, have open access to the valuable content created through publicly funded research. In support of this, repository networks work together to provide seamless access to research outputs and adopt common practices that maximize the ethical re-use of content and development of value added services."<sup>4</sup>

Interoperability is a crucial underlying requirement for realizing this vision. As stated in the State of Repository Interoperability Report, "the potential to create a unified body of scholarly materials is entirely reliant on interoperability – specifically, that repositories follow consistent guidelines, protocols, and standards for interoperability which allow them to communicate with each other; connect with other systems; and transfer information, metadata, and digital objects between each other."<sup>5</sup>

<sup>3</sup> Ibid pg. 4

<sup>4</sup> Kathleen Shearer. Towards a Seamless Global Research Infrastructure: Report of the Aligning Repository Networks Meeting, March 2014. pg. 6 Available at: <https://www.coar-repositories.org/files/Aligning-Repository-Networks-Meeting-Report.pdf>

<sup>5</sup> Ibid

For the purposes of this report, the scope of the topic of interoperability has been interpreted in a very broad sense in order to avoid overlooking relevant topics. There was no strict limitation on technical aspects in order to cover all facets of technically driven interoperability of the repository stakeholders. In addition, the meaning of interoperability includes methods and approaches to exposing content and metadata within a given system.

The goal of this roadmap is to define the interoperability cornerstones for repositories according to their relevance and level of complexity, with particular attention paid to the following challenges:

### 1. Technical

- implications for APIs, metadata formats, added-value services, vocabularies
- linking to other entities (publications, research data, project information, impact / statistics)
- sharing of digital assets
- coverage and support of digital assets beyond text

### 2. Organizational

- roles and responsibilities of operation, support, development

### 3. Legal

- issues on data exchange and re-use

The objectives are to provide a detailed account of repository interoperability issues in order to:

- Describe all areas of interoperability for repositories in the future, including:
  - researchers needs and workflows (creating, reading, re-using, discovery and filtering as well as extracting of knowledge; front and backend),
  - funders and institutional interests,
  - integration with other infrastructures (e.g. disciplinary research infrastructures, authority services; other (incl. commercial) stakeholders (e.g. Google Scholar, Mendeley, ResearchGate, F1000));
- Identify levels of complexity, timing and importance for each key area of interoperability;
- Develop a list of priority issues for interoperability efforts for the repository community.

## 2.2 User requirements

As noted earlier in the report, repositories have evolved to play a variety of roles, such as providing open access to content, contributing to research evaluation systems, publishing platforms, and so on. Below is a comprehensive list of the user requirements for repositories based on different stakeholder communities:

Stakeholder	Users requirements
Researcher as an author	<ul style="list-style-type: none"> <li>• Easy metadata feeds (including re-using existing data)</li> <li>• Upload documents easily</li> <li>• Easy and comfortable creation of complex data relations</li> <li>• Automatic addition of linked data</li> <li>• High visibility of his digital objects/documents/scientific profile and relations</li> <li>• Easy embedding of publications in different working environments (personal publication lists, virtual research environments, etc.)</li> <li>• Comfortable creating of complex documents (enhanced publications)</li> <li>• Transparent usage statistics (download and citation frequencies)</li> <li>• Easy storage and publishing solutions for articles, journals, monographs, working papers</li> </ul>
Researcher as reader/end user	<ul style="list-style-type: none"> <li>• Open Access to publications</li> <li>• Visible references of their publications in secondary environments</li> <li>• Comfortable search tools</li> <li>• Visualized complex information on publication relationships (to other (similar or recommended) publications, to related research data)</li> <li>• Transparent bibliometric information</li> <li>• Stable document links</li> <li>• Stable and safe document storage (Long-term preservation)</li> </ul>
Institution	<ul style="list-style-type: none"> <li>• Exposure of their affiliated publication output (institutional bibliography)</li> <li>• Exposure of related institutional research information (projects, prizes etc.)</li> <li>• Document and report research output information for assessment and compliance monitoring</li> </ul>
Funder	<ul style="list-style-type: none"> <li>• Assess impact of funded research outcome</li> <li>• Provide open access to research outputs</li> <li>• Track and monitor research outputs</li> </ul>
External stakeholder ( <i>publisher, information company, service provider</i> )	<ul style="list-style-type: none"> <li>• Comprehensive, high quality, and standardized metadata information on publications and research data in order to reuse them</li> </ul>

## 2.3 Participating systems and stakeholders

Currently there is an increasing number of systems dealing with academic information supply (this includes producing, managing and referencing publications). The major systems are described below:

- Aggregator Services
- Bibliographic Management Tools
- Current Research Information Systems (CRISs)
- Digital Collections
- Discipline-based Repositories
- E-Learning Systems
- Hosting Services
- Internet Search Engines
- Local Library Systems (catalogues)
- Publication Management Systems
- Publishing Systems (journals, monographs)
- Research Data Repositories
- Virtual Research Environments (VREs)
- Other Global Services and Players

All of these systems deal in some way with either full text publications and/or other types of research outputs, or contain information about publications. As such, bi-directional interoperability channels for the exchange of information with repository systems are obvious. In particular, we are seeing the emergence of numerous approaches for providing and fetching metadata that allow new pathways for linking documents and objects with each other. These new forms of linking data have the power to overcome the role of data silos and to support a new quality of metadata scope and related services.

**Aggregator Services** collect and combine metadata (and sometimes full text) from numerous repositories and expose the aggregated contents in different interfaces (end-user orientated, OAI-PMH based). Because of the normalization needs to integrate the different sub-systems the specific aspects of their interoperability behaviour have to be considered.

**Current Research Information Systems (CRIS)** manage and track information about the research process. They collect a wide range of information and metadata about all aspects of the research activity carried out at an institution. Since publications play an important role in this context, these systems can be integrated with repository functions or to communicate with independent repository systems. This leads to specific interoperability needs with institutional repositories.

**Digital Collections** are often hosted and managed by institutions that also maintain open access repositories. The content includes additional material such as image collections, digitized books, historic maps, digitized historical scientific publications, source material. Scientific publications may also be included.

**Discipline-based Repositories** have been established by many scientific communities to share publications. Interoperability between these systems and institution based systems is needed for exchanging data (in order to re-use them) and for avoiding redundant efforts. In some disciplines, researchers primarily use subject repositories, as with ArXiV/INSPIRE or PubMed. Thus, interoperability between institutional and subject repositories is paramount, if uptake of the institutional repository by local researchers is the goal. This interoperability can go both ways, importing bibliographic data from the subject repository to the IR but also importing full text, e.g. 'green' versions' or dissertations from IR to subject repositories.

**E-Learning Systems** support the scholarly communication and then produce their own documents or learning objects (those can typically be found uploaded in repositories). Usually these systems integrate publications into their course material (as citations and uploads).

**Hosting Services (for repositories)** are available as non-commercial and commercial systems which host institutional repositories. Usually those systems deliver a common set of interoperability channels which include specific ways of communication to the local repository environments.

**Internet Search Engines** are the entry point for most users of the content in repositories. Especially important is that repository collections are exposed and indexable by search engines.

**Local Library Systems** are a source of bibliographic data for publications metadata. In addition, some repository systems also provide an interface to cataloging systems with the option to push new repository entries into the cataloging module. In this context, bi-directional metadata exchange with the institutional repository is highly desirable.

**Publication Management Systems** aim to manage and reference the publications of institutions and their researchers and thus have a different objective related to publications than repositories. The focus of these systems is on collecting bibliographic references and not the full text documents. There is a trend whereby some academic institutions are extending the scope of these systems to and adding in the functionalities to the existing IR system or integrating the two systems into one broader platform (University Gent/Belgium, University Lund/Sweden). Wherever possible, there should be interoperability between publication management systems and full text open access repositories.

**Publishing Systems** (in particular Open Journals System for electronic journals and its descendants for monographs and conference proceedings) are widely used to produce journals, monographs or conference proceedings, typically in an academic environment and by the same institution which maintains institutional repositories. In addition, since the authors may be affiliated with the local academic institution the output is (or should be) included in the corresponding institutional repository.

**Research Data Repositories** also can be hosted and managed by the same institution or publications in a repository may point to the related dataset housed in a data repository. Interoperability between these systems will be crucial for linking related content.

**Virtual Research Environments (VREs)** are platforms for collaborative research activities and will probably play a strong role in future developments in the research process. Since this approach includes storage of publications and research data management activities there are overlapping aspects with repository functionality.

**Global Services and Players** including platforms that provide researchers CVs and bios, such as ResearchGate, Mendeley, and Google Scholar play a relevant role. Aspects such as the visibility of these services or using and importing their data to avoid duplicate efforts make interoperability highly desirable.

### 3 Interoperability Issues

The diverse roles of repositories and their relationships with other existing or evolving systems predicate the need for interoperability across numerous interfaces and identities. However, it is not necessarily clear in which of the many areas of interoperability the repository community should focus its efforts. To that end, the COAR Interoperability Working Group launched a process to assist in identifying priorities for interoperability work, and to raise awareness of the issues of interoperability with the broader community. They developed a comprehensive list of 6 interoperability topics, divided into 35 issues.

<p><b>Impact and Visibility</b></p> <ul style="list-style-type: none"> <li>• Supporting Search Engine Optimization (SEO)</li> <li>• Supporting Repository Ranking Systems</li> <li>• Exposing Usage Statistics</li> <li>• Exposing Bibliometric Information</li> <li>• Supporting Visibility in Repository Registries</li> <li>• Improving Registry Infrastructure</li> </ul> <p><b>Usability</b></p> <ul style="list-style-type: none"> <li>• Supporting Authorization and Authentication</li> <li>• Supporting Embedding Services</li> <li>• Exposing Publication Lists</li> <li>• Exposing Citation Formats</li> <li>• Supporting Data Export Functions</li> <li>• Integrating Availability Services</li> <li>• Supporting Author Identification Systems</li> <li>• Supporting Institutional Services</li> <li>• Extending End-User Usability</li> <li>• Extending Usage of Visualization Tools</li> </ul> <p><b>Sustainability</b></p> <ul style="list-style-type: none"> <li>• Improving Platform Stability</li> <li>• Supporting Long-term Preservation and Archiving</li> <li>• Exposing Persistent Identifiers</li> <li>• Integrating different Persistent Identifiers</li> </ul>	<p><b>Data Issues</b></p> <ul style="list-style-type: none"> <li>• Supporting additional Metadata Format(s)</li> <li>• Improving Metadata Quality (Data Curation)</li> <li>• Supporting Enhanced Publications</li> <li>• Supporting Linked (Open) Data</li> <li>• Publication of Research Data</li> <li>• Handling of Complex/Compound/Nested Repository Objects</li> <li>• Monitoring Open Access Mandate Compliance</li> <li>• Exposing Versioning Information</li> </ul> <p><b>Validation and Aggregation</b></p> <ul style="list-style-type: none"> <li>• Validating Repository Metadata</li> <li>• Processing Related Full-text</li> <li>• De Duplication</li> </ul> <p><b>Technical Issues</b></p> <ul style="list-style-type: none"> <li>• Defining Architectural Recommendations for Repositories and their Interoperability</li> <li>• Extending/Replacing Metadata Exposition Protocols</li> <li>• Supporting OAI Service Provider Usage</li> <li>• Supporting Deposit Protocols</li> </ul>
---	---

## 4 Results and Analysis

This list of issues was shared with an Expert Advisory Panel, which prioritized them in terms of their perceived importance and complexity. Based on their feedback, the issues have been analyzed according to their perceived impact in three areas: general relevance, timing and complexity. General relevance was determined based on how many times the issue was identified by experts. Time factor was determined by experts choosing immediate, medium or long-term relevance and complexity was determined by experts assigning either low, moderate or high levels of complexity. The full results including experts' comments are available in Appendix I.

In order to better identify priorities, the issues have been ranked according to strategic need, immediate relevance and low complexity. Comparing the aspects in the list deliver a priority-sorted and can be read as priority-based roadmap for implementing the issues. Merging and weighting the different aspects requires a more flexible representation, which visualizes the priorities. Graphs that visualize the ranking of each issue is available in the Appendix.

### 4.1 Priorities according to topic area

The six high level topics are shown below, organized according to their scores by expert reviewers.

#### General relevance

High	Moderate	Low
<ul style="list-style-type: none"> <li>• Usability</li> <li>• Data Issues</li> </ul>	<ul style="list-style-type: none"> <li>• Impact and Visibility</li> </ul>	<ul style="list-style-type: none"> <li>• Validation and Aggregation</li> <li>• Sustainability</li> <li>• Technical Issues</li> </ul>

#### Timing

Immediate	Medium-term	Future
<ul style="list-style-type: none"> <li>• Impact and Visibility</li> <li>• Sustainability</li> </ul>	<ul style="list-style-type: none"> <li>• Validation and Aggregation</li> <li>• Data Issues</li> <li>• Usability</li> </ul>	<ul style="list-style-type: none"> <li>• Technical Issues</li> </ul>

## Complexity

Low	Moderate	High
	<ul style="list-style-type: none"> <li>• Technical Issues</li> <li>• Validation and Aggregation</li> <li>• Usability</li> <li>• Sustainability</li> <li>• Impact and Visibility</li> </ul>	<ul style="list-style-type: none"> <li>• Data Issues</li> </ul>

## 4.2 Priorities according to specific issues

The same approach has been used to rank the specific interoperability issues that fall under each topic.

### General Relevance

This table lists the issues, in descending order for the relevance, and categorized according to high, moderate and low relevance.

High	Moderate	Low
<ul style="list-style-type: none"> <li>• Exposing Usage Statistics</li> <li>• Supporting Enhanced Publications</li> <li>• Supporting Long-term Preservation and Archiving</li> <li>• Improving Metadata Quality (Data Curation)</li> <li>• Improving Registry Infrastructure</li> <li>• Supporting additional Metadata format(s)</li> <li>• Exposing Citation Formats</li> <li>• Exposing Persistent Identifiers</li> <li>• Publication of Research Data</li> <li>• Supporting Embedding Services</li> <li>• Supporting Linked (Open) Data</li> <li>• Supporting Visibility in Repository Registries</li> <li>• Exposing Bibliometric Information</li> <li>• Supporting Author Identification Systems</li> <li>• Validating Repository Metadata</li> </ul>	<ul style="list-style-type: none"> <li>• Extending Usage of Visualization Tools</li> <li>• Handling of complex / compound / nested Repository Objects</li> <li>• Exposing Versioning Information</li> <li>• Extending End-User Usability</li> <li>• Supporting Search Engine Optimization (SEO)</li> <li>• De Duplication</li> <li>• Extending/Replacing Metadata Exposition Protocols</li> </ul>	<ul style="list-style-type: none"> <li>• Processing Related Fulltext</li> <li>• Defining Architectural Recommendations for Repositories and their Interoperability</li> <li>• Supporting Authorization and Authentication</li> <li>• Supporting Repository Ranking Systems</li> <li>• Supporting OAI Service Provider Usage</li> <li>• Integrating Availability Services</li> <li>• Improving Platform Stability</li> <li>• Integrating different Persistent Identifiers</li> <li>• Supporting Deposit Protocols</li> <li>• Supporting Institutional Services</li> </ul>

<ul style="list-style-type: none"> <li>• Exposing Publication Lists</li> <li>• Supporting Data Export functions</li> <li>• Monitoring Open Access Mandate Compliance</li> </ul>		
---	--	--

## Complexity

Low	Moderate	High
<ul style="list-style-type: none"> <li>• Supporting Visibility in Repository Registries</li> <li>• Supporting Search Engine Optimization (SEO)</li> <li>• Exposing Publication Lists</li> <li>• Exposing Citation Formats</li> <li>• Supporting Data Export functions</li> <li>• Exposing Persistent Identifiers</li> <li>• Supporting Author Identification Systems</li> <li>• Integrating Availability Services</li> <li>• Supporting Embedding Services</li> <li>• Validating Repository Metadata</li> <li>• Improving Platform Stability</li> <li>• Supporting Authorization and Authentication</li> <li>• Supporting Institutional Services</li> <li>• Integrating different Persistent Identifiers</li> <li>• Extending End-User Usability</li> <li>• Monitoring Open Access Mandate Compliance</li> </ul>	<ul style="list-style-type: none"> <li>• Exposing Versioning Information</li> <li>• Improving Registry Infrastructure</li> <li>• Supporting Repository Ranking Systems</li> <li>• Supporting OAI Service Provider Usage</li> <li>• Exposing Bibliometric Information</li> <li>• De Duplication</li> <li>• Exposing Usage Statistics</li> <li>• Supporting additional Metadata format(s)</li> </ul>	<ul style="list-style-type: none"> <li>• Extending Usage of Visualization Tools</li> <li>• Publication of Research Data</li> <li>• Defining Architectural Recommendations for Repositories and their Interoperability</li> <li>• Supporting Deposit Protocols</li> <li>• Extending/Replacing Metadata Exposition Protocols</li> <li>• Processing Related Fulltext</li> <li>• Supporting Linked (Open) Data</li> <li>• Improving Metadata Quality (Data Curation)</li> <li>• Supporting Enhanced Publications</li> <li>• Supporting Long-term Preservation and Archiving</li> <li>• Handling of Complex/Compound/Nested Repository Objects</li> </ul>

## Temporal relevance and complexity

The following list combines two major aspects, timing with complexity as second dimension.

	<b>Short term</b>	<b>Medium term</b>	<b>Long term</b>
<b>Low Complexity</b>	<ul style="list-style-type: none"> <li>• Exposing Citation Formats</li> <li>• Supporting Data Export Functions</li> <li>• Supporting Author Identification Systems</li> <li>• Supporting Search Engine Optimization (SEO)</li> <li>• Exposing Publication Lists</li> <li>• Integrating Different Persistent Identifiers</li> </ul>	<ul style="list-style-type: none"> <li>• Exposing Persistent Identifiers</li> <li>• Supporting Authorization and Authentication</li> <li>• Improving Platform Stability</li> <li>• Supporting Institutional Services</li> <li>• Extending End-User Usability</li> <li>• Validating Repository Metadata</li> <li>• Supporting Visibility in Repository Registries</li> <li>• Supporting OAI Service Provider Usage</li> <li>• Integrating Availability Services</li> <li>• Supporting Embedding Services</li> <li>• Supporting Repository Ranking Systems</li> </ul>	
<b>Moderate Complexity</b>	<ul style="list-style-type: none"> <li>• Exposing Bibliometric Information</li> </ul>	<ul style="list-style-type: none"> <li>• Exposing Versioning Information</li> <li>• De Duplication</li> <li>• Improving Registry Infrastructure</li> <li>• Monitoring Open Access Mandate Compliance</li> </ul>	
<b>High Complexity</b>	<ul style="list-style-type: none"> <li>• Exposing Usage Statistics</li> <li>• Supporting Additional Metadata Format(s)</li> </ul>	<ul style="list-style-type: none"> <li>• Publication of Research Data</li> <li>• Improving Metadata Quality (Data Curation)</li> <li>• Processing Related Full Text</li> <li>• Supporting Deposit Protocols</li> <li>• Defining Architectural Recommendations for Repositories and their Interoperability</li> <li>• Supporting Enhanced Publications</li> </ul>	<ul style="list-style-type: none"> <li>• Extending Usage of Visualization Tools</li> <li>• Supporting Linked (Open) Data</li> <li>• Extending/Replacing Metadata Exposition Protocols</li> <li>• Handling of Complex/Compound/Nested Repository Objects</li> <li>• Supporting Long-term Preservation and Archiving</li> </ul>

## Immediate Priorities

Based on the results, 9 issues fall into the immediate timing, with high and moderate general relevance. These issues are considered, therefore, highest priority for interoperability work.

1. Exposing citation formats
2. Supporting data export functions
3. Supporting author identification systems
4. Exposing publication lists
5. Exposing bibliometric information
6. Exposing usage statistics
7. Supporting additional metadata format(s)

Two other issues were determined to have immediate priority and moderate general relevance:

8. Integrating different persistent identifiers
9. Supporting search engine optimization (SEO)

These issues are described in more detail along with selected comments from the experts.

- 1. Exposing citation formats:** Bibliographic metadata are cited in various formats. In this context repositories may expose their metadata in those citation formats. There are technical implementations available to integrate the creation of numerous popular citation formats.

*“Should be standard since it is one of the few real services that a repository could possibly offer”*

- 2. Supporting data export functions:** Export functionality delivers the metadata in order to process them elsewhere later and thus contains a broader range of metadata information than citations only.

*“Should be standard since it is one of the few real services that a repository could possibly offer”*

- 3. Supporting author identification systems:** Author identification systems intend to identify academic authors in a unique way. Different systems are available (ORCID, ResearcherID, AuthorClaim etc.) and can be used in the repository context.

*“IMHO, a very important extension of current repositories, but relying on external authority data maintained by external systems / staff; needs a lookup service for persons, which can be applied during indexing and searching of data”... “Clearly supports open systems like ORCID instead of proprietary ones”*

- 4. Exposing publication lists:** Personal and institutional publication lists can offer a service of institutional repositories based on the stored bibliographic metadata and relations with people and institutions.

*“Normally, these publications only contain a subset of all publications (from persons or institutions), hence it will be a crucial question how to enrich and integrate these lists with complementary data.”... “Very important”*

5. **Exposing bibliometric information:** Bibliometric methods provide information about citation frequencies of repository objects. Repositories can use and provide this information as extended metadata, for example displaying supplementary information on the publication landing page.

*“Complexity lies in transparency and aggregation” .... “Needs to be a centralized (funded?) effort.”*

6. **Exposing usage statistics:** Usage statistics deliver figures about document usage in repositories. Statistical tools exist to analyse log-files of individual repository instances. These data provide an informative basis about end-users utilization. However a number of challenges need to be solved to make usage statistics comparable on a global level, e.g. agreements on robot filter list, methods for log data normalization, prove of usage statistics indicators etc.

*“The basic standards and guidelines are available; proof of concept has been shown by initiatives like OA-Stats, SURFSure, IRUS-UK, OpenAIRE; It’s missing its large scale roll-out and agreements on an organisational level”*

7. **Supporting additional metadata format(s):** Currently repositories deliver metadata mostly via OAI-PMH in Dublin Core format as mandatory and some of them support a broad variety of extended formats. Since DC is interrelated with a limited number of tags and a certain vagueness of interpretation there is a strong need to agree for alternative, more convenient metadata formats offering finer granularity. Potential formats to be considered (and depending on the purpose) are MODS, METS, MARC, CERIF and others.

*“Adding more standards when they bring richness and detail is a key step to move forward in the current situation. The complexity of course depends on number and complexity of the new adoptions. DC is no longer useful for advancing in the field.”*

8. **Integrating different persistent identifiers:** Repositories do often provide their own identifier system, which is applied to the objects which are generated through this repository. With the increasing exchange (import/export/harvesting/LOD) of metadata, different identifier systems will ‘clash’ in a repository environment.

*“Any kind of additional, persistent identifier is useful to fetch more, contextual? Information, e.g. PMID/PMCID, ISSN, DOI”*

9. **Supporting Search Engine Optimization (SEO):** SEO methods are focused on optimizing the ranking of web sites and their contents in search engines. Activities to optimize the ranking of web sites are known as Web Boosting.

*“There can be some common guidelines provided (layout issues, HTML headers, sitemaps, advice how to get into Google, using social networks like Twitter to speed up indexing etc.)”*

## 5 Conclusion

The results of the roadmap process demonstrate that exchange and re-use of metadata from repositories represent a key priority for interoperability work. This will support cross-repository and cross-system interoperability (for example with CRIS and research biographical systems). It also speaks to the critical importance for the repository community in terms of integrating with other systems in order to provide standardized information for the purposes of research administrative systems. To achieve this, standardized vocabularies, metadata schemas and elements will be key.

COAR is already initiated work on several of these issues:

**COAR-CASRAI Working Group** is developing a blueprint that will outline the steps needed to ensure greater interoperability across repository networks, and ideally with other related systems and actors as well (e.g. CRIS). The group has members from COAR and CASRAI, EuroCRIS, Jisc, La Referencia, OpenAIRE and SHARE (ARL). The group began its work in Nov 2014 via a mix of teleconference and online collaboration tools and the aim to have a completed plan by March 2015.

**COAR Controlled Vocabularies for Repository Assets Interest Group** is updating and maintaining a vocabulary developed in the European context, and widen its applicability for global use. Updates are currently underway and will be made publicly available as a resource for the global repository community. The vocabulary will be maintained via an international governance structure using established workflows.

**COAR Interest Group on Usage Data and Beyond** is undertaking a review of existing and emerging article level metrics with the aim of identifying a common set of metrics that can be adopted across repositories, enabling the community to compare statistics across repositories. There are already a number of projects at the local, national and international level that are gathering and aggregating usage data from repositories.

**COAR Impact and Visibility Interest Group** is exploring and documenting different approaches for maximizing repository visibility as well as develop new and innovative strategies for adoption by organizations around the world. Among other things, the group will be looking at implementing linked data into repositories as well as best practices for search engine optimization.

Each of these groups is looking at some aspect of interoperability for repositories. Still, there is much more work to be done. For example, adoption of standards by individual repositories is always challenging and requires significant outreach with repository managers. This is even more difficult if we look at global interoperability, across regions which have tremendous diversity in terms of languages, requirements, and perspectives.

One very promising way of ensuring widespread adoption will be to have them implemented directly into repository software platforms. After which the local repository manager can activate the new features, especially by initiating the version upgrade and related configuration activities.

In order to enforce interoperability in a bi-directional scope repository managers and software developments should pay particular attention to ensuring their interfaces are open to work with other systems.

COAR will consider various paths for improving interoperability in the priority areas:

- What work is involved in ensuring interoperability in priority areas?
- Which stakeholders must be included in implementation and how can we best engage them in these activities? Particularly important will be the participation of the repository platform developers, as this is an essential strategy for widespread adoption.

In terms of next steps, COAR will:

1. Disseminate the roadmap and its results to COAR members and the broader community of stakeholders, in particular:
  - a. Regional/National Repository Networks
  - b. Repository Platform Communities
  - c. Repository Managers
  - d. Other related stakeholders (e.g. research administrative communities, publishers)
2. Build support and awareness of the benefits and need for interoperability
3. Support dialogue and progress towards the adoption of common approaches across regions and stakeholder communities
4. Develop and undertake strategies for implementing standards in repositories

Clearly, as a global organization, COAR has an important role to play in connecting these various communities and coalescing around some best practices. In addition, COAR can coordinate the essential efforts for preparing underlying definitions, recommendations and guidelines to assist the development and implementation process.

## Appendix I: The Glossary

### ***Author Identification Systems***

Author Identification systems intend to identify and disambiguate academic authors. Different systems are available (ORCID, ResearcherID, AuthorClaim etc.) and can be used in the repository context.

### ***Authorization and Authentification***

Services to handle authorization and authentification (especially in an institutional background) in order to control access and role support.

### ***Availability Services***

Link resolving services calculate and deliver information about the availability of objects, thus this service can be relevant for repositories in a bi-directional approach, e.g. adding availability links and playing a role as a target system for included resources.

### ***Bibliometrics***

Bibliometric methods provide information about citation frequencies of repository objects. Repositories can use and provide this information as extended metadata for the specific objects, for example displaying supplementary information on their publication landing page.

### ***Citation Formats***

Publications are cited in various well-defined and frequently used formats (Harvard, Chicago, APA, MLA and numerous others). In this context repositories may expose their bibliographic metadata in those citation formats. There are technical implementations available (for example the Citation Style Engine using the Citation Style Language, CSL) to integrate the creation and display of numerous popular citation formats based on the bibliographic metadata stored in repositories.

### ***Data Curation***

Metadata quality improvements based on data curation activities (both automatically and human-driven) have extensive consequences for the quality of data-reuse. In this sense observing Guidelines for metadata requirements (OpenAIRE Guidelines (extending the former DRIVER Guidelines), DINI Certificate, RIOXX) and improving the repositories compliance will have a strong impact on metadata quality.

### ***Deduplication***

Deduplication includes methods to recognize and identify identical or similar objects (duplicates and versions) and to prepare storing and providing this information through the repository.

### ***Digital Collections***

Digital Collections are related with digital libraries and digitization projects. The content includes material as image collections, digitized books, historic maps, digitized historical publications and different source material. Scientific publications are included often as well. These platforms are in many cases hosted and managed by the same institution as institutional repositories, provide an OAI-PMH interface normally and the hosted material is typically open access.

### ***Enhanced Publications***

Enhanced publications are digital objects enriched with corresponding material as video, audio, images, research data, project information, citations/references, comments/annotations and thus will play a more and more prominent role in future publishing and research activities.

### ***Linked (Open) Data***

In the semantic web context linked open data and their provision play a relevant role in future interoperable data processing activities and thus can become a future service of repositories. Vice versa repositories could extend and provide their metadata through using and merging external semantic data.

### ***Long-term Preservation and Archiving***

Long-term preservation includes efforts to ensure continued access to digital materials (regardless of technology changes over time while archiving includes access services to the material).

### ***Metadata format(s)***

Currently repositories deliver metadata mostly via OAI-PMH with Dublin Core as mandatory format and many of them support a broad variety of extended formats additionally. Since DC is interrelated with a limited number of tags and a certain vagueness of interpretation there is a strong need to agree for alternative, more convenient metadata formats offering finer granularity and better metadata quality. Potential formats to be considered (and depending on the purpose) are MODS, METS, MARC, CERIF and others. In order to take into account how additional metadata formats could be processed efficiently the current distribution of delivered metadata formats is relevant.

### ***Publication Lists***

Listing the publications of scientists or institutions is a key functionality of repositories with publication management services. To provide interfaces to embed those lists in external interfaces is a obvious additional interoperability channel for repositories.

### ***Repository Ranking Systems***

Automatic checking and validating repositories including the processing of ranking scores are standard services. Additionally by help of cybermetric methods the web representation of sites can be measured in order to describe their relevance. To this respect the “Consejo Superior de Investigaciones Científicas” (CSIC) has developed the “Ranking web of repositories” to compare “the web presence and the web impact (link visibility) of their contents”. Supporting strategies to optimize the results of these systems can be helpful.

## **Repository Registries**

Repository registries list repositories and their characteristics under different aspects (national, disciplines, platform technology, publication type) and provide information for the repository community and to external stakeholders about them and their interoperability status.

## **Research Data**

The publication and re-use of research data is a increasing field of scholarly communication and therefore relevant for institutional repositories to handle this type of data. Solutions how to handle and integrate this data into the repository have to be designed and developed.

## **Search Engine Optimization (SEO)**

SEO methods are focused on optimizing the ranking of web sites and their contents in search engines. Activities to optimize the ranking of web sites are known as Web Boosting. These strategies can play an important role in improving the visibility of repositories and their documents.

## **Usage Statistics**

Usage statistics deliver figures about document usage in repositories. Statistical tools exist to analyse log-files of individual repository instances. These data provide an informative basis about end-users utilization. However a number of challenges need to be solved to make usage statistics comparable on a global level, e.g. agreements on robot filter lists, methods for log data normalization, prove of usage statistics indicators etc.

## **Versioning Information**

Repositories include similar versions of documents. In order to make this transparent methods to identify and expose the corresponding information have to be developed and applied.

## **Visualization Tools**

Repositories contain complex data structures of metadata, contents and additional information and especially interconnections between them and external resources at an increasing rate. Visualization tools may help to make this information more transparent and understandable and accordingly may optimize end-user interoperability.

## Appendix 2: Acronyms and Abbreviations

agINFRA	Agricultural Data Infrastructure
AP	Application Profile
APA	American Psychological Association (citation style)
BASE	Bielefeld Academic Search Engine
CASRAI	Consortia Advancing Standards in Research Administration Information
CERIF	Common European Research Information Format
CKAN	Comprehensive Knowledge Archive Network
COAR	Confederation of Open Access Repositories
CRIS	Current Research Information System
CSIC	Consejo Superior de Investigaciones Científicas
CSL	Citation Style Language
DC	Dublin Core Metadata Standard
DINI	Deutsche Initiative für Netzwerkinformation
DOI	Digital Object Identifier
DRIVER	Digital Repository Infrastructure Vision for European Research
EDM	European Data Model
EUDAT	European Data Infrastructure
IPR	Intellectual Property Rights
IR	Institutional Repository
ISSN	International Standard Serial Number
LOD	Linked Open Data
LTP	Long-term Preservation
MARC	Machine-Readable Cataloging
MLA	Modern Language Association (citation style)
METS	Metadata Encoding and Transmission Standard
MODS	Metadata Object Description Schema
NBN	National Bibliographic Number
NIH	National Institutes of Health, United States
OA	Open Access
OAI-ORE	Open Archives Initiative – Object Reuse and Exchange
OAI-PMH	Open Archives Initiative – Protocol for Metadata Harvesting
OAR	Open Access Repositories
OAS	OA-Statistik
OpenAIRE	Open Access Infrastructure Research for Europe
OpenDOAR	Directory of Open Access Repositories
ORCID	Open Researcher & Contributor ID
PI	Persistent Identifier
PID	Persistent Identifier
PMC	PubMed Central
PMID	PubMed Indexing Number
PMCID	PubMed Central Referencing Number
RDF	Resource Description Framework
REST	Representational State Transfer
RIOXX	Repository Interoperability Opportunities Extension
SEO	Search engine optimization
SHARE	SHared Access Research Ecosystem
SPARQL	SPARQL Protocol And RDF Query Language
SURE	Statistics on the Usage of Repositories
SWORD	Simple Web-service Offering Repository Deposit
URL	Uniform Resource Locators
URN	Uniform Resource Names
VRE	Virtual Research Environment

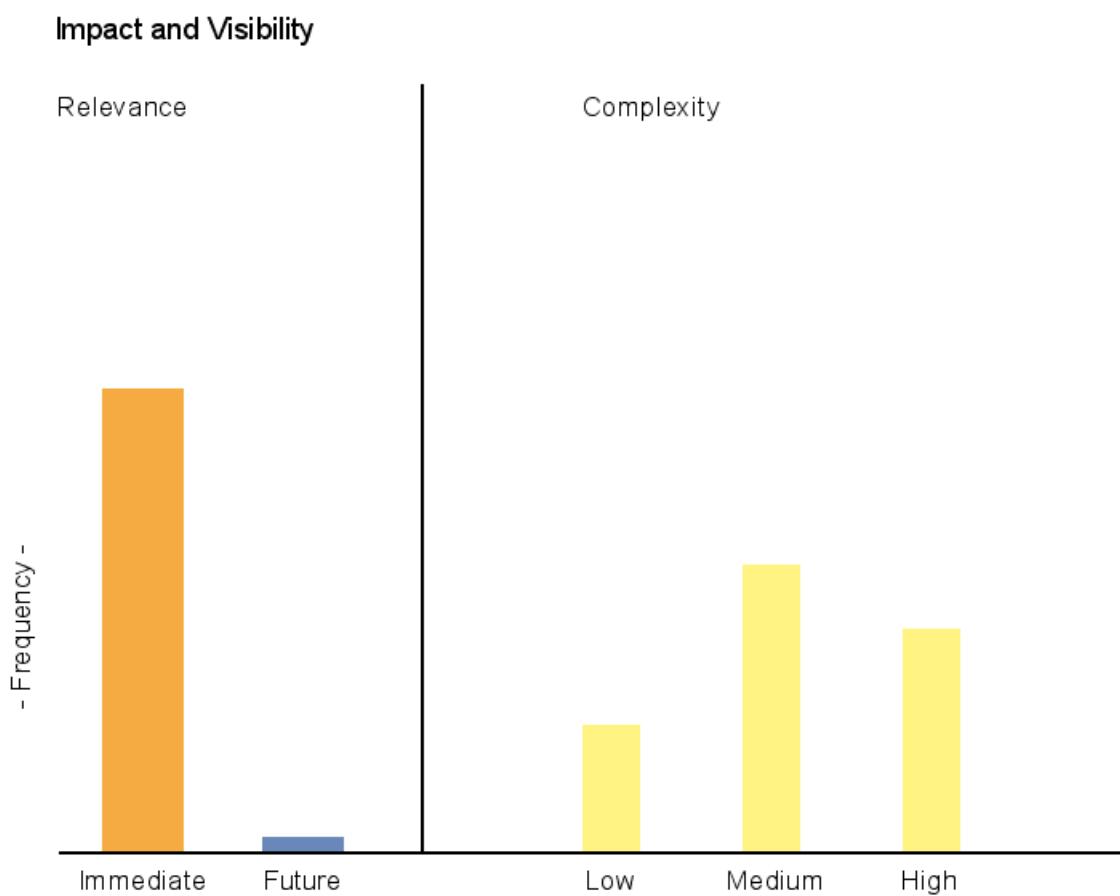
## Appendix 3: The questionnaire and its response

The feedback text has been processed automatically in order to derive and compare the relevance and complexity level of the single topics. The following bar diagrams show the corresponding specific values and a selective list of valuable comments from the feedback texts.

### Key Aspect: Impact and Visibility

#### Strategic Benefit:

Supporting the visibility of repositories and their contents including their relevance, usage and impact measures:



## Comments

*“Open Access is not any longer the domain of repositories only. Publishers align their business models; some CRIS claim to support bibliographic management even of OA publications.”*

*“I think the strategic issue is to work on the visibility/impact of the content held in our repositories; repositories themselves should be understood as infrastructure, seamlessly integrated in the wider research infrastructure and so, become eventually invisible.”*

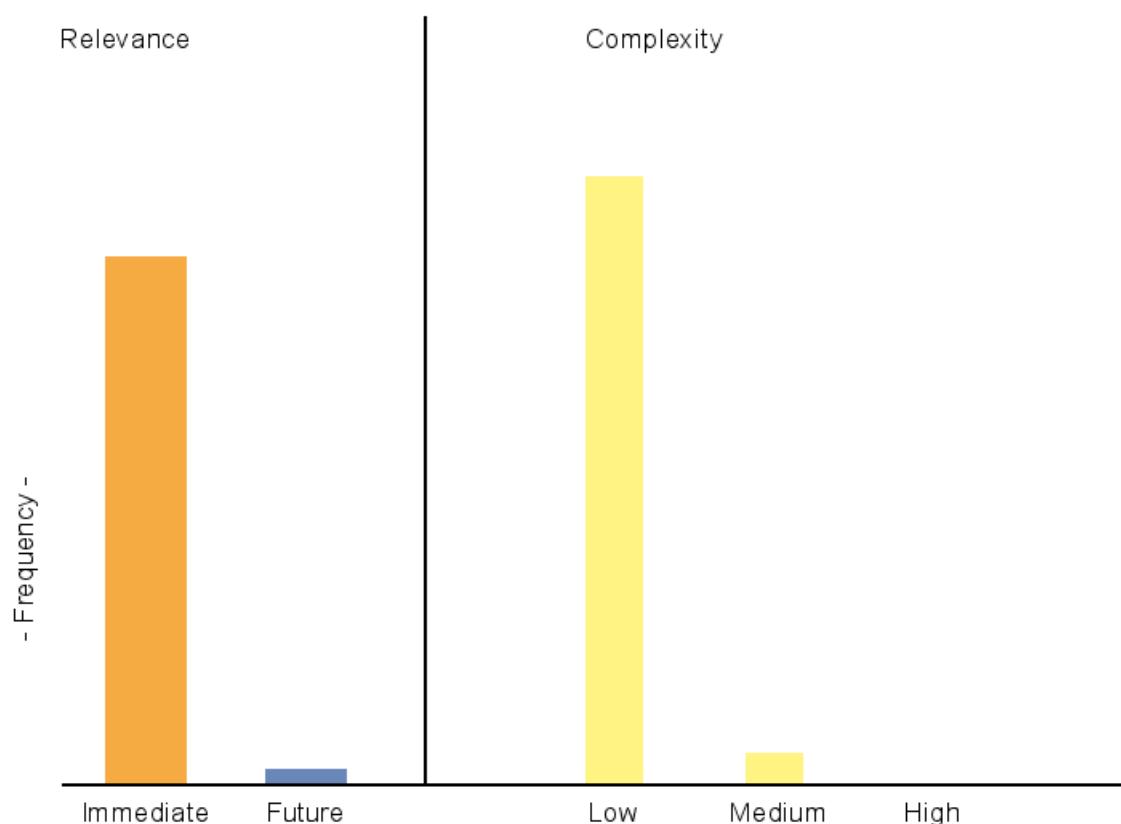
*“Visibility should especially be supported in the context of popular search engines (e.g., by optimizing crawling by Google Scholar) and social media channels (e.g., by automatically tweeting new entries or stats under @myrepos).”*

*“Impact and visibility are most important issues from a researcher’s point of view.”*

## Issue: Supporting Search Engine Optimization (SEO)

SEO methods are focused on optimizing the ranking of web sites and their contents in search engines. Activities to optimize the ranking of web sites are known as Web Boosting.

### Supporting Search Engine Optimization (SEO)



## Comments

*“If this is a system to make the repository website well positioned in google, I think it is just against repositories content visibility; individual repositories websites (and isolated content in the repository) are not important, and we do not want to get into a repositories website competition.”*

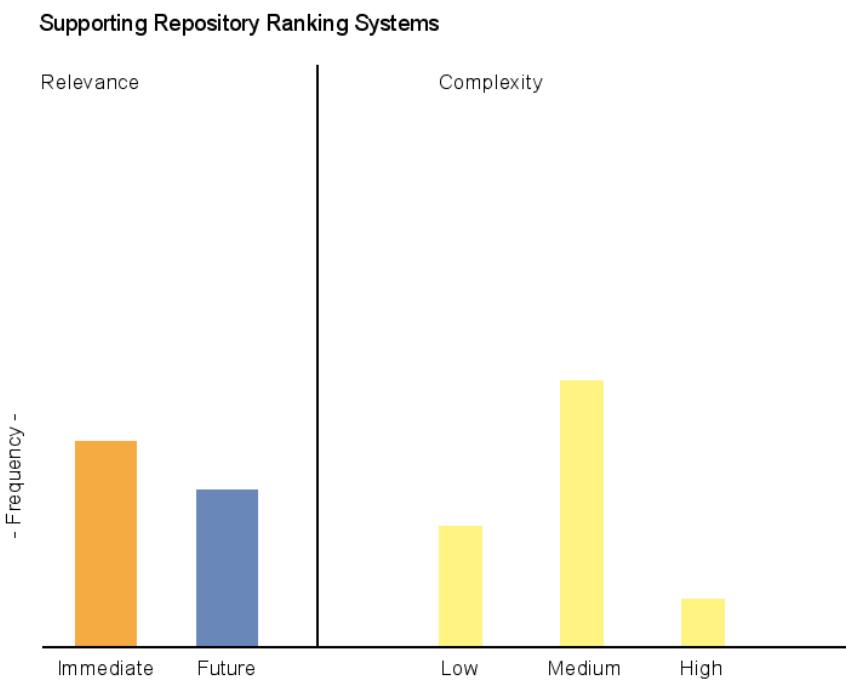
*“There can be some common guidelines provided (layout issues, HTML headers, sitemaps, advice how to get into Google, using social networks like Twitter to speed up indexing etc.)”*

*“Should be no more than some best practices /guidelines.”*

*“Complexity is low, but requires specialized staff oriented towards marketing, this may be a barrier in public organizations.”*

## Issue: Supporting Repository Ranking Systems

Automatic checking and validating repositories including the processing of ranking scores are standard services. Additionally by help of cybermetric methods the web representation of sites can be measured in order to describe their relevance. To this respect the “Consejo Superior de Investigaciones Científicas” (CSIC) has developed the “Ranking web of repositories” to compare “the **web presence** and the **web impact** (link visibility) of their contents”. Supporting strategies to optimize the results of these systems can be helpful.



## Comments

*“Other rankings that take the content of the repository and its services into account should be developed.”*

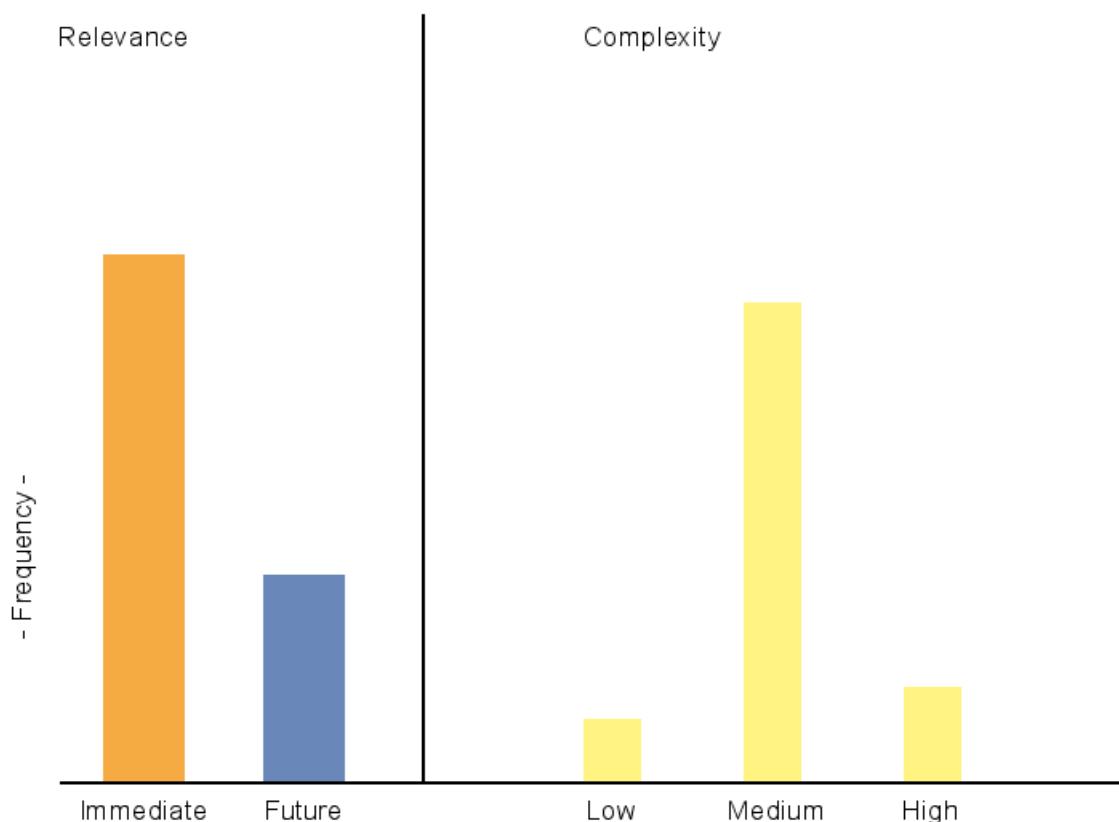
*“Technically is not difficult, but coming up with objective metrics and rankings would require a lot of elaboration, as the ‘quality’ of repositories can be seen from many perspectives. CSIC metric is related to Web presence, but this is only one aspect.”*

*“This is more an issue of transparency and acceptance than of implementation.”*

## Issue: Exposing Usage Statistics

Usage statistics deliver figures about document usage in repositories. Statistical tools exist to analyse log-files of individual repository instances. These data provide an informative basis about end-users utilization. However a number of challenges need to be solved to make usage statistics comparable on a global level, e.g. agreements on robot filter list, methods for log data normalization, prove of usage statistics indicators etc.

### Exposing Usage Statistics



## Comments

*“Usage statistics need standards and if that comes in, it gets difficult to implement such services, see OAS.”*

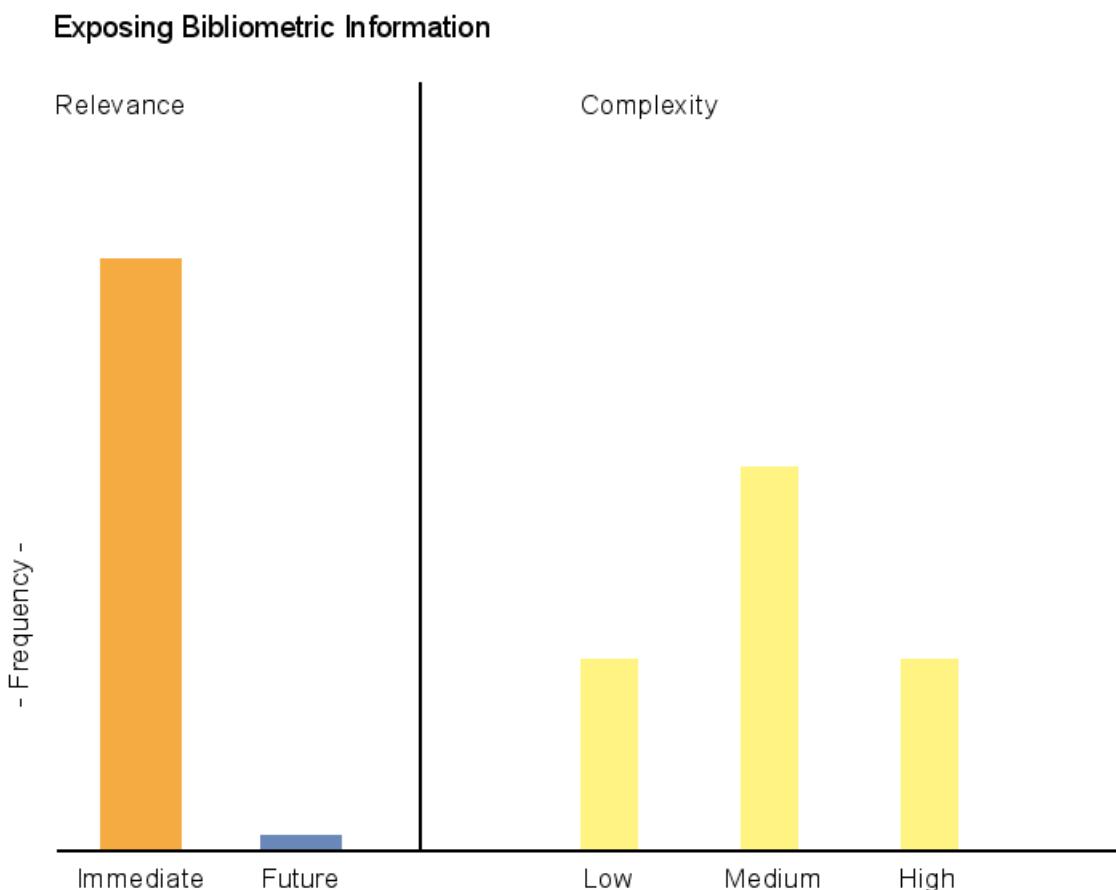
*“I think this issue is very relevant for repositories to offer alternatives to traditional quality and impact metrics”*

*“The basic standards and guidelines are available; proof of concept has been shown by initiatives like OA-Stats, SURFSure, IRUS-UK, OpenAIRE; It’s missing its large scale roll-out and agreements on an organisational level”*

*“Log data normalization to provide meaningful results is very hard to create and tune and willingness to expose this is getting less.”*

## Issue: Exposing Bibliometric Information

Bibliometric methods provide information about citation frequencies of repository objects. Repositories can use and provide this information as extended metadata, for example displaying supplementary information on the publication landing page.



## Comments

“Complexity lies in transparency and aggregation.”

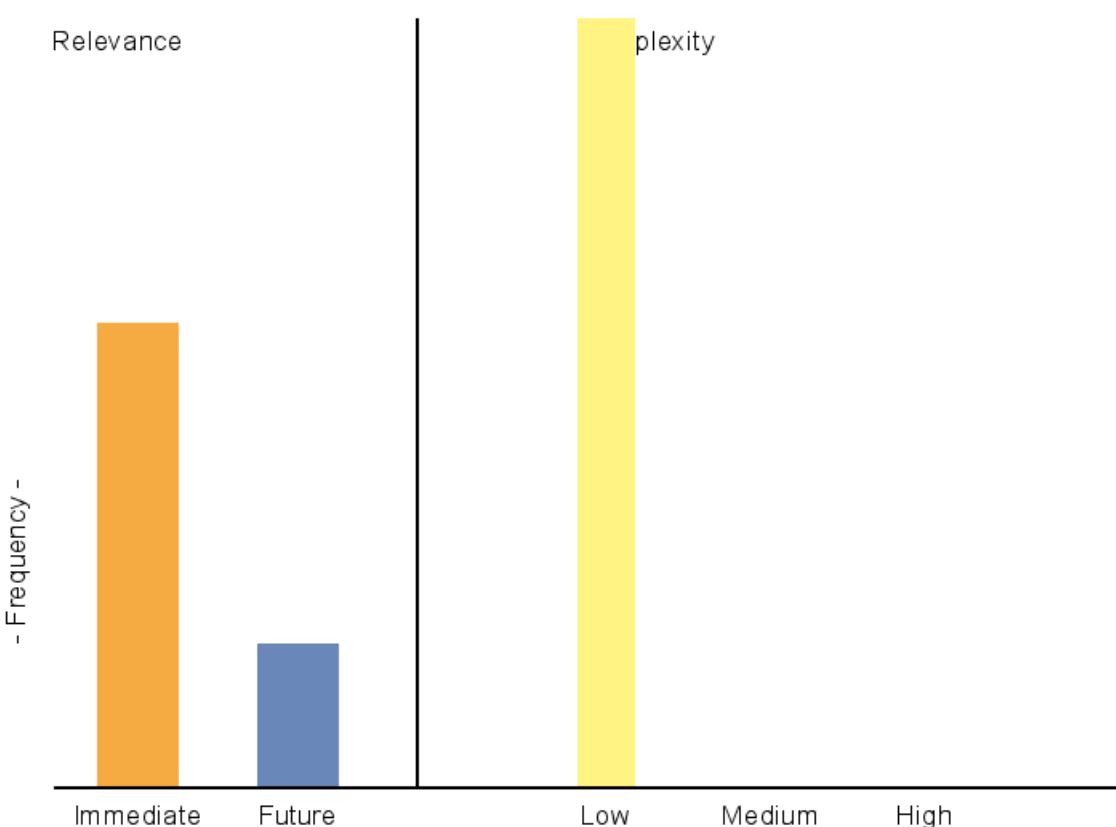
“Needs to be a centralized (funded?) effort.”

“Technically is not a problem. The problem is what kind of bibliometrics can be used at the repository level. We know them at personal, institutional or even country level. Maybe the institutional level is the one we should address here?”

## Issue: Supporting Visibility in Repository Registries

Repository registries list repositories and their characteristics under different aspects (national, disciplines, platform technology, publication type) and provide information to external stakeholders about them and their interoperability status.

### Supporting Visibility in Repository Registries



## Comments

*“Since most registries are not up to date and cover only basic information, registries in contrast to a more comprehensive ranking, are obsolete.”*

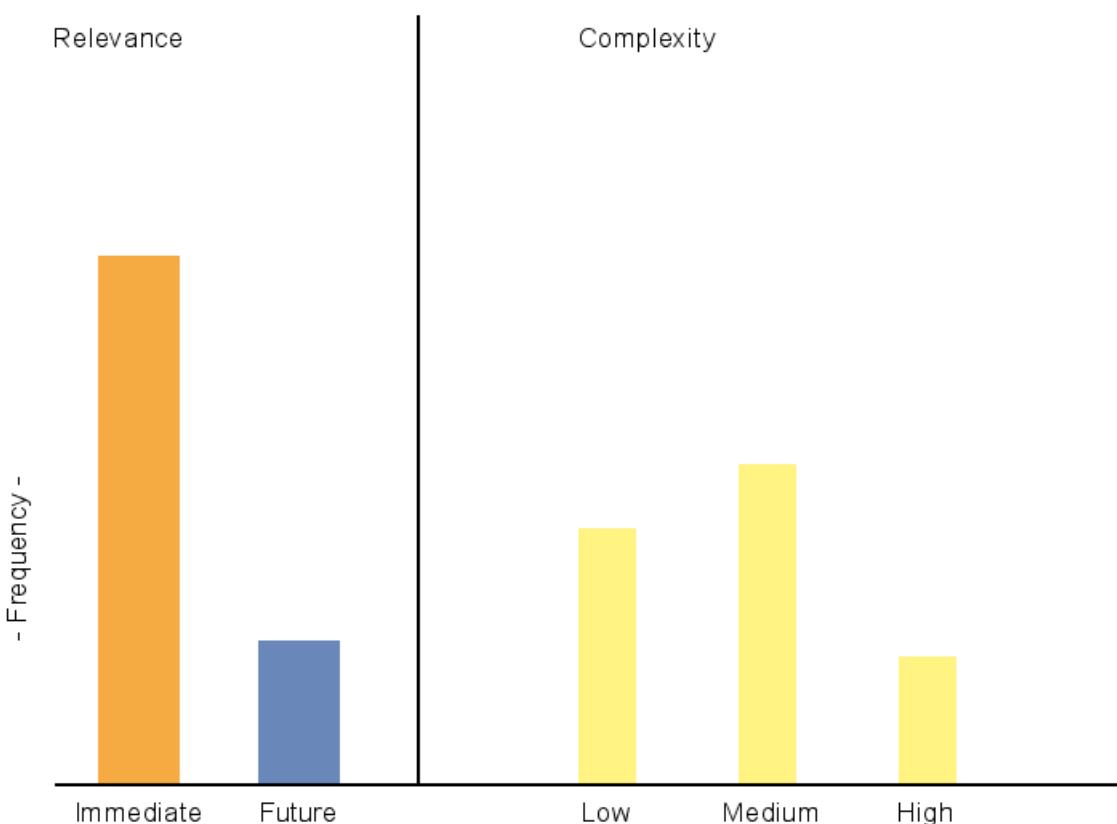
*“This is easy to do and would add a lot of benefit to users.”*

*“Primarily to provide deposit service.”*

## Issue: Improving Registry Infrastructure

Current implementations of registries lack specific information for interoperability purposes, especially a unique and persistent repository identifier and information about classifications used, vocabularies supported, access policy, full text percentage, content analysis. This information will be very valuable for the quality of communication processes with interested stakeholders.

### Improving Registry Infrastructure



## Comments

*“Additional information could be delivered via the description element of the OAI Identify command.”*

*“Complexity lies in the heterogeneous local repository infrastructure.”*

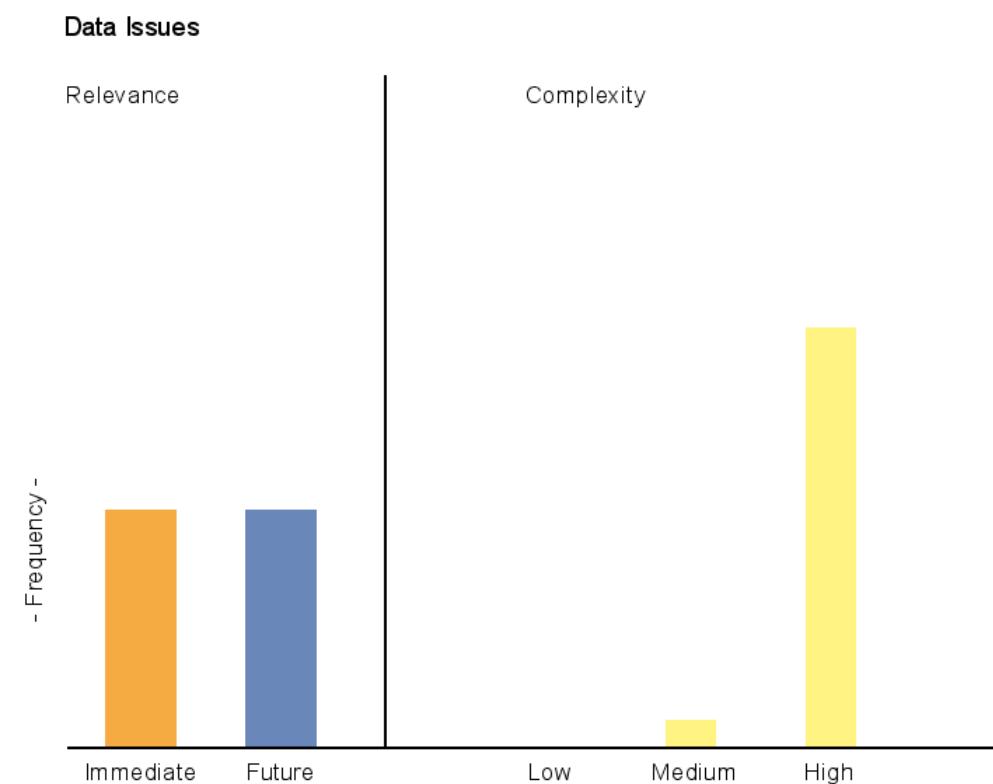
*“Develop description set profiles and application profiles (AP). Apply or improve tools that generate APs automatically.”*

## Key Aspect: Data Issues

### Strategic Benefit:

Rethinking the unit and form of the scholarly publication as a complex, heterogeneous, dynamic object and methods to handle them in a repository environment.

The process of scholarly communication in the digital era is driven by technology, open access policies, a culture of open science and the contextualization of research output. As a result e.g. new forms of publication types are attempted and need to be supported in future. More cooperative approaches based on modern researcher communication will produce new forms of publications (for example “liquid publications”).



## Comments

*“This is a central element for the future of repositories. However, it is difficult to achieve because it requires a shift in technology and what is more important, a reconsideration of current standards for describing “Research Objects”.”*

*“Absolutely critical in a data-driven open science framework”*

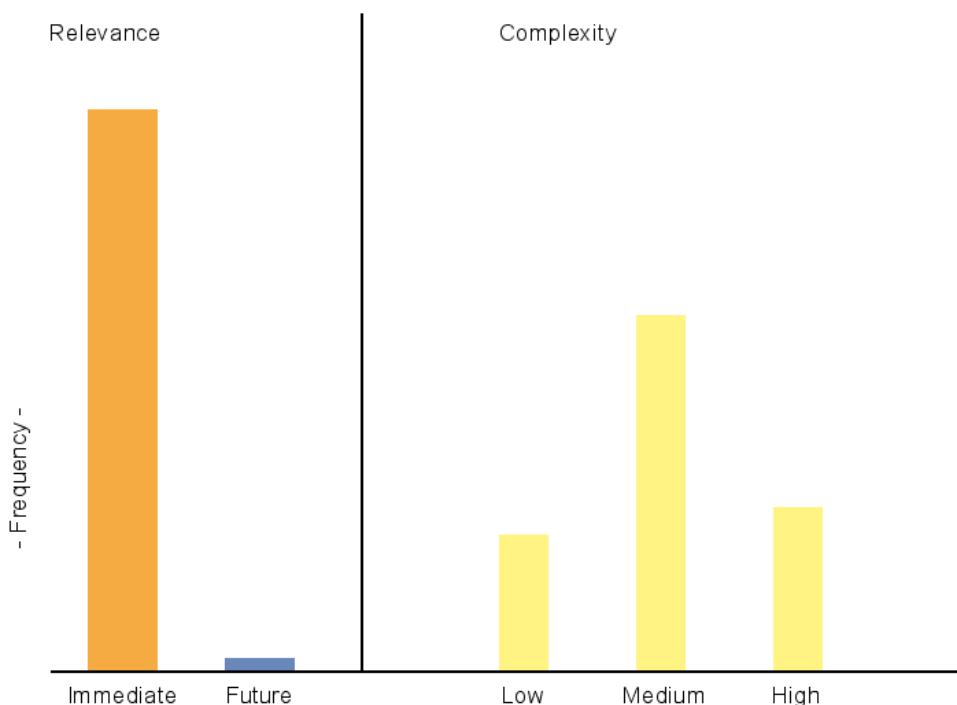
*“Central issue but currently no scalable or generic solution available”*

*“I do not know if repositories really should go in that direction. One of their current benefits might be the relatively monolithic, straightforward presentation of research material. Only with respect to research data I see some urgent need to opening up repositories for this category.”*

## Issue: Supporting additional Metadata format(s)

Currently repositories deliver metadata mostly via OAI-PMH in Dublin Core format as mandatory and some of them support a broad variety of extended formats. Since DC is interrelated with a limited number of tags and a certain vagueness of interpretation there is a strong need to agree for alternative, more convenient metadata formats offering finer granularity. Potential formats to be considered (and depending on the purpose) are MODS, METS, MARC, CERIF and others.

Supporting additional Metadata format(s)



## Comments

*“Broader discussion among repo stakeholders, guidelines and training needed”*

*“If DC as a generic format is not good enough, then it needs to be improved or replaced. We don’t want additional formats for the same purpose.”*

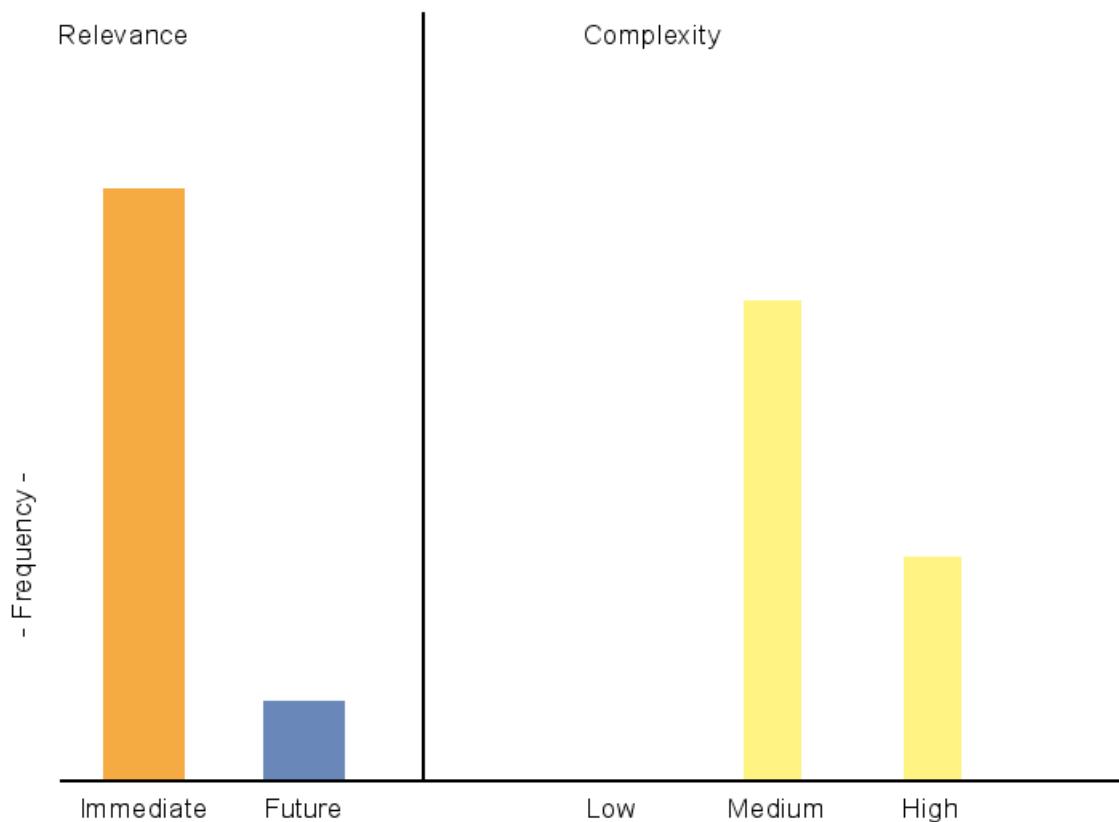
*“Depends on community and complexity of additional format”*

*“Adding more standards when they bring richness and detail is a key step to move forward in the current situation. The complexity of course depends on number and complexity of the new adoptions. DC is no longer useful for advancing in the field.”*

## Issue: Improving Metadata Quality (Data Curation)

Metadata quality improvements based on data curation activities (both automatically and human-driven) have extensive effects for the quality of data-reuse. In this sense observing Guidelines for metadata requirements (OpenAIRE Guidelines (extending former DRIVER Guidelines), DINI Certificate, RIOXX) and improving the repositories compliance will have impact on metadata quality.

### Improving Metadata Quality (Data Curation)



## Comments

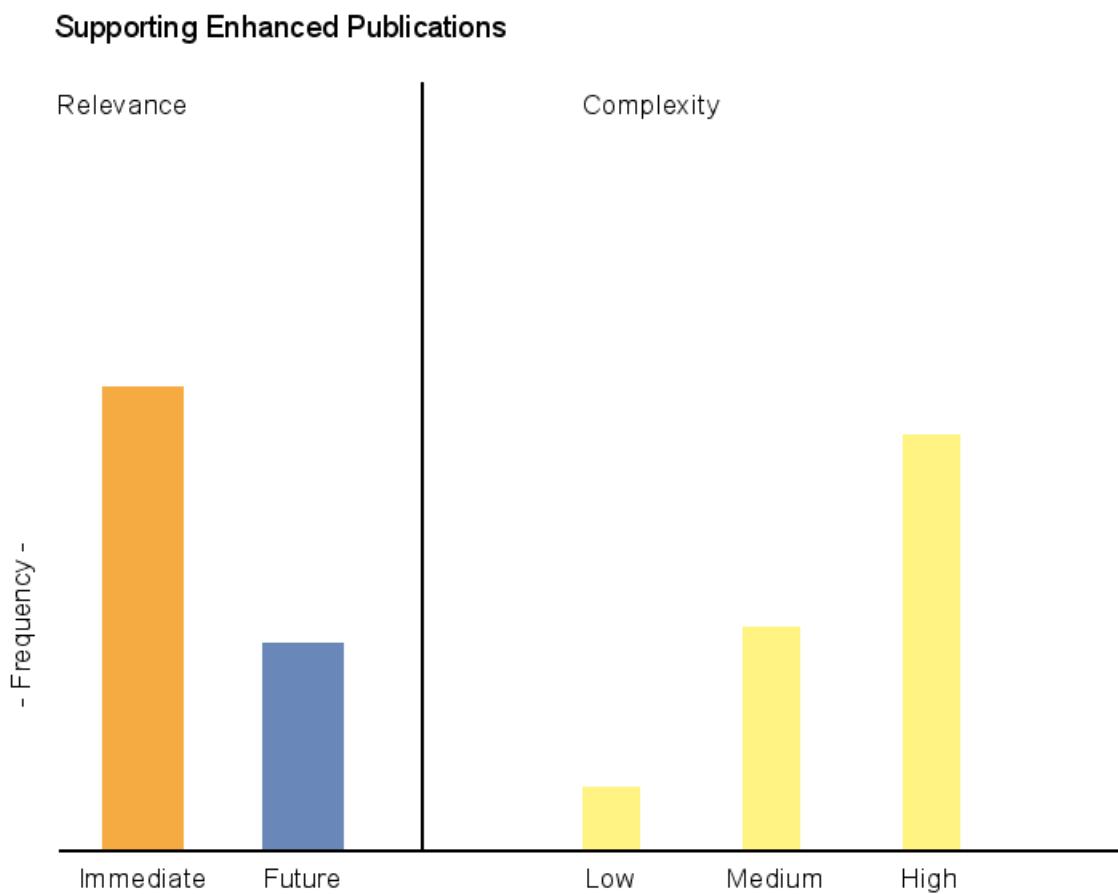
*“It is the Achilles’ heel of repositories; the high standard of metadata quality is set by commercial citation databases”*

*“It's not complex but it's hard to achieve.”*

*“This is important but will probably only occur slowly, as tools and quality methods are better understood and managed.”*

## Issue: Supporting Enhanced Publications

Enhanced publications are enriched digital objects with corresponding and additional material as video, audio, research data, project information, citations/references, comments/annotations and thus will play a more and more prominent role in publishing and research activities.



## Comments

*“Enhanced publications are an important step in the development of OAR.”*

*“High because it requires one organization to increase the complexity of their work or because it requires collaboration between multiple organizations. Also because these complex objects need to be persistent.”*

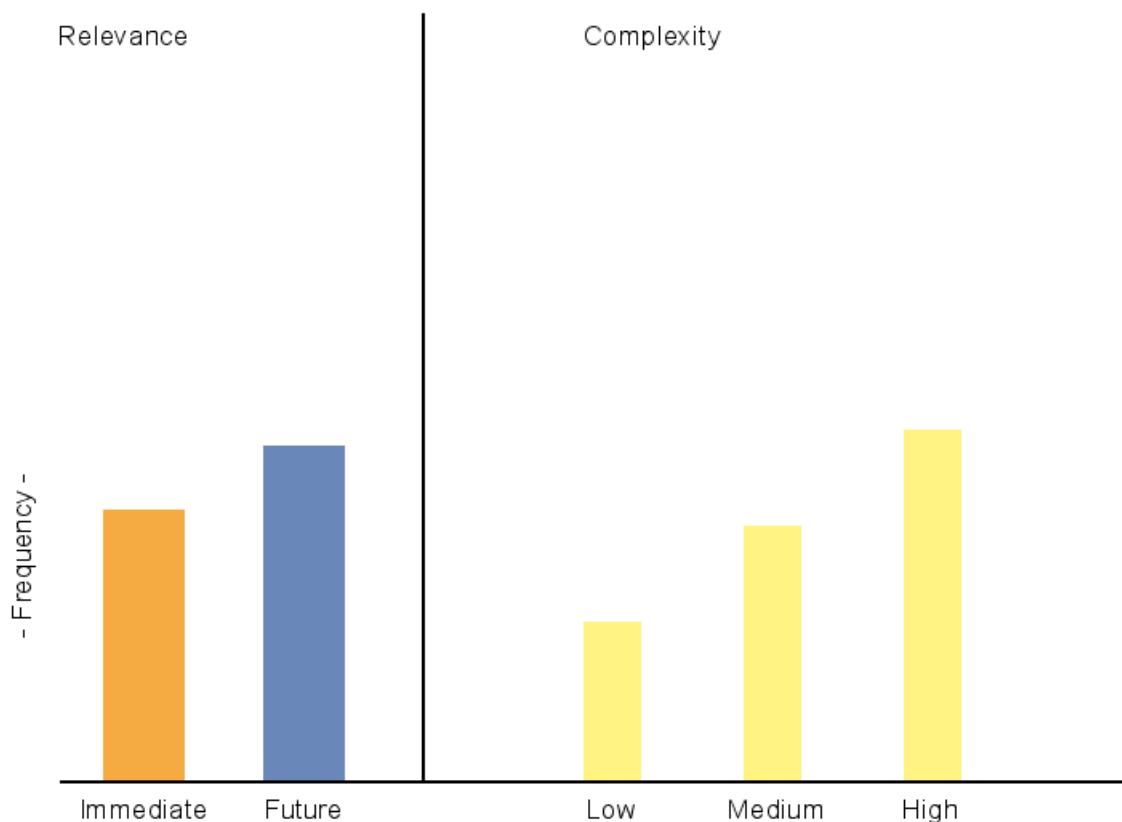
*“Current workflows were not typically built to deal with these ever evolving objects with loosely coupled relations.”*

*“Non-sensical without user services that take advantage of these capabilities”*

## Issue: Supporting Linked (Open) Data

In the semantic web context linked open data and their provision play a relevant role in future interoperable data processing and thus can become a future service of repositories. Vice versa repositories could extend and provide their metadata through using and merging external semantic data.

### Supporting Linked (Open) Data



## Comments

*“In my opinion only stepping into the semantic web repositories will remain relevant to support research”*

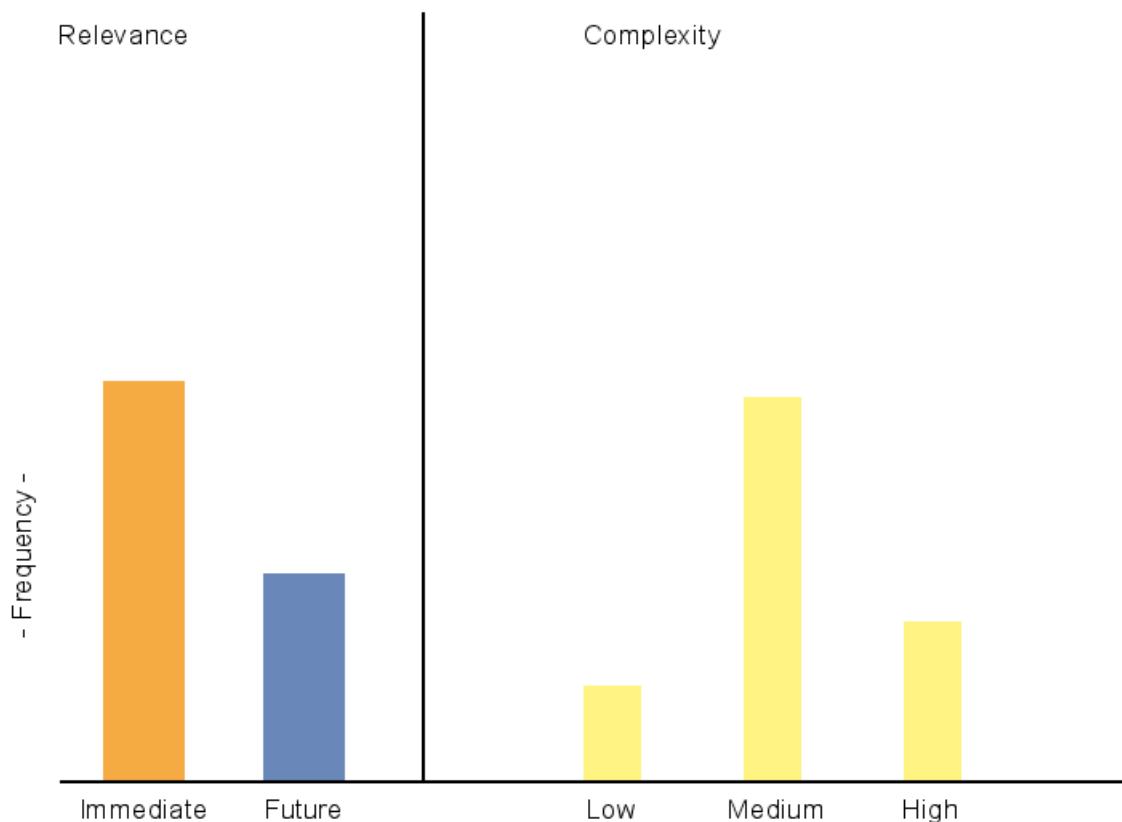
*“Publishing a repository’s content (at least the metadata) as LOD can become important for third party applications to mash-up this data with other’s data. It would provide a good alternative to OAI-PMH, if current infrastructure issues of synchronizing LOD are solved. Since publishing as LOD in any case means interlinking the data with external sources by means of typed relations, it would foster the topic of data interoperability. LOD will definitely increase the requirements towards system administration, e.g. by maintaining a SPARQL endpoint.”*

*“Future of LOD as a universal concept still unclear”*

## Issue: Publication of Research Data

The publication and re-use of research data is a increasing field of scholarly communication and therefore relevant for institutional repositories to handle this type of data. Solutions how to handle this data have to be designed and developed.

### Publication of Research Data



## Comments

*“I don’t see this as the domain of repositories storing scholarly publications; of course hybrid systems are possible; references to research data should be made if applicable.”*

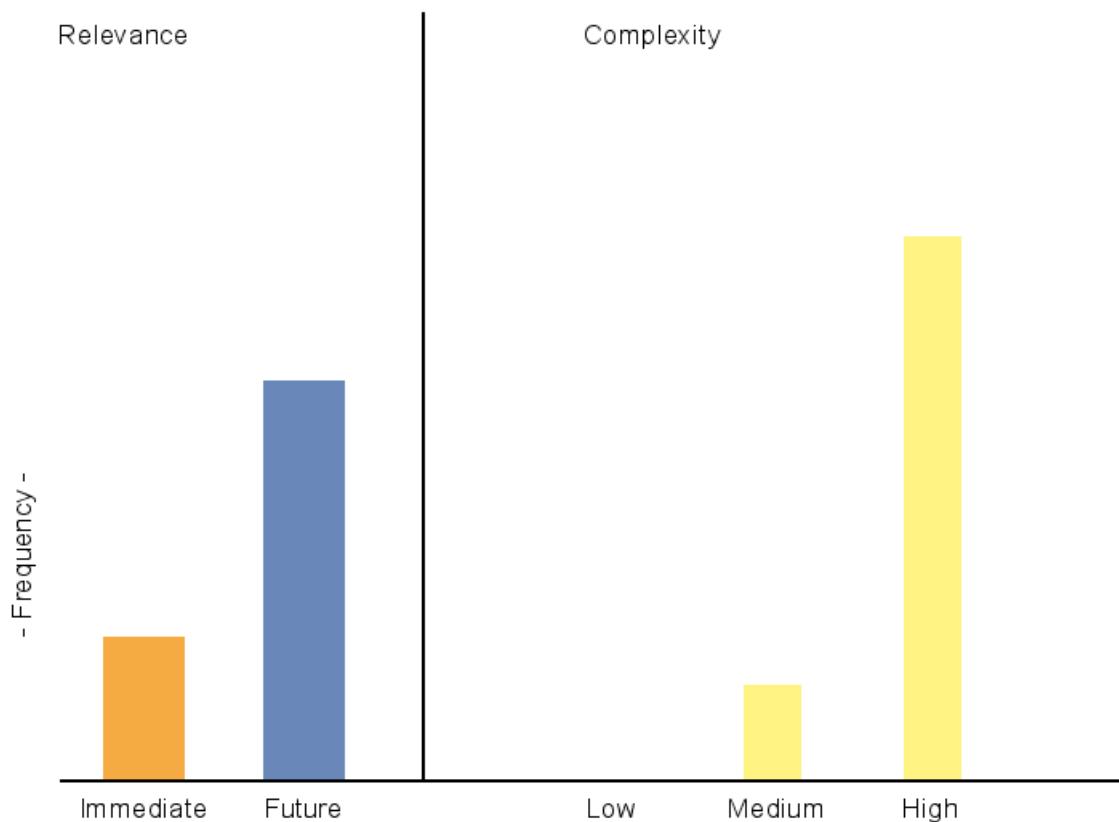
*“Research data should not be managed by repositories themselves, but by appropriate Research data platforms like e.g. CKAN. The integration with the repository can be quite straightforward, but also cumbersome if the repository is regarded as the leading system.”*

*“There are several solutions for this, either in house or relying on external systems. In any case, this is basically a problem of creating a culture of sharing, so the problem does not lie at the technical level.”*

## Issue: Handling of complex/compound/nested Repository Objects

Traditional repositories are focused on monolithic resources, often allowing just one repository object to be recorded and delivered. Traditional catalog systems provide nested structures, e.g. for capturing volumes and their included items and this approach could be valuable in the repository context too.

### Handling of complex/compound/nested Repository Objects



## Comments

*“This is rather a question of the internal data model and its flexibility.”*

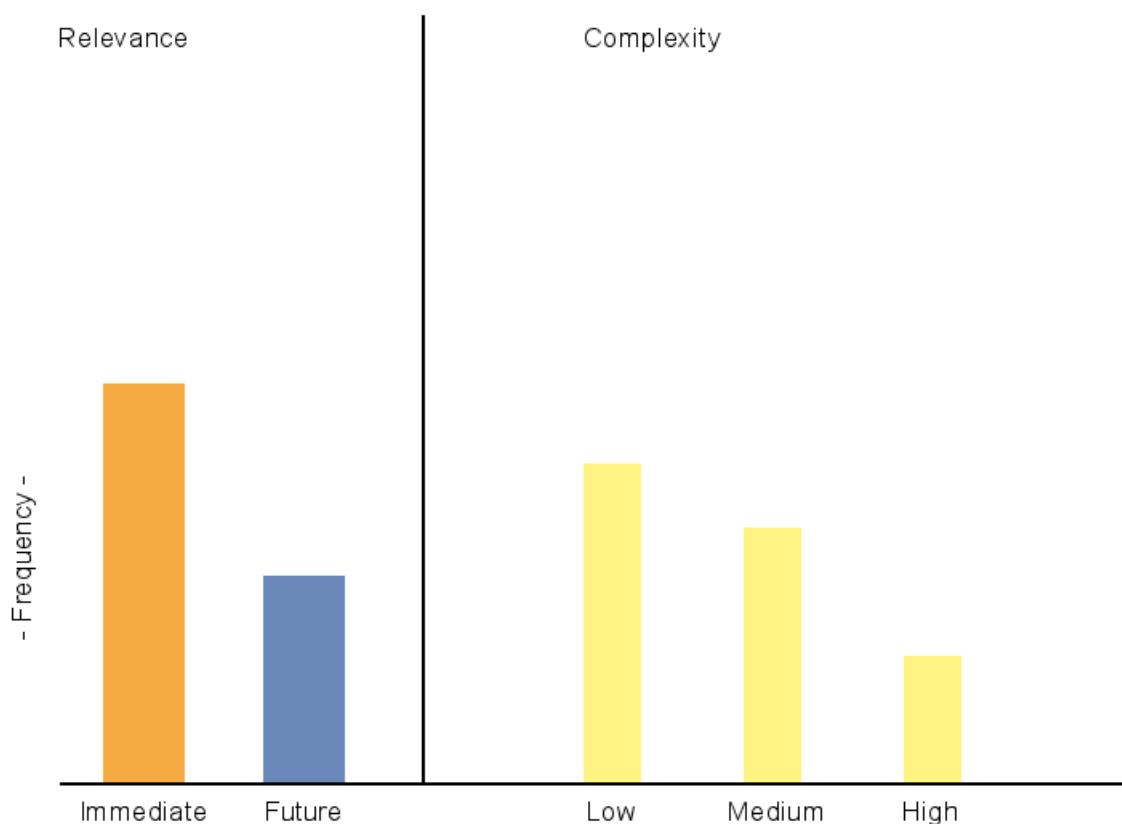
*“The complex object that repositories will have to handle is a complex, distributed, global relationships ‘virtual object’ that go beyond single repositories. Linked data seems to me again, the way forward for repositories”*

*“Although OAI-ORE has been issued a long time ago, this can still be a very complex task, depending on the document types we handle. I’m in favour of adapting/reusing the maximum number of features from the EDM in order to maintain interoperability with Europeana and other future applications built on it.”*

## Issue: Monitoring Open Access Mandate Compliance

Tools to control repositories and their contents regarding their compliance with mandates can be helpful for research, funders and institutions.

### Monitoring Open Access Mandate Compliance



## Comments

*“IMHO, a ‘simple’ (which is actually not that simple dealing with all these response codes) link checker would do it.. OA policy could also be encoded in the description element of Identify call.”*

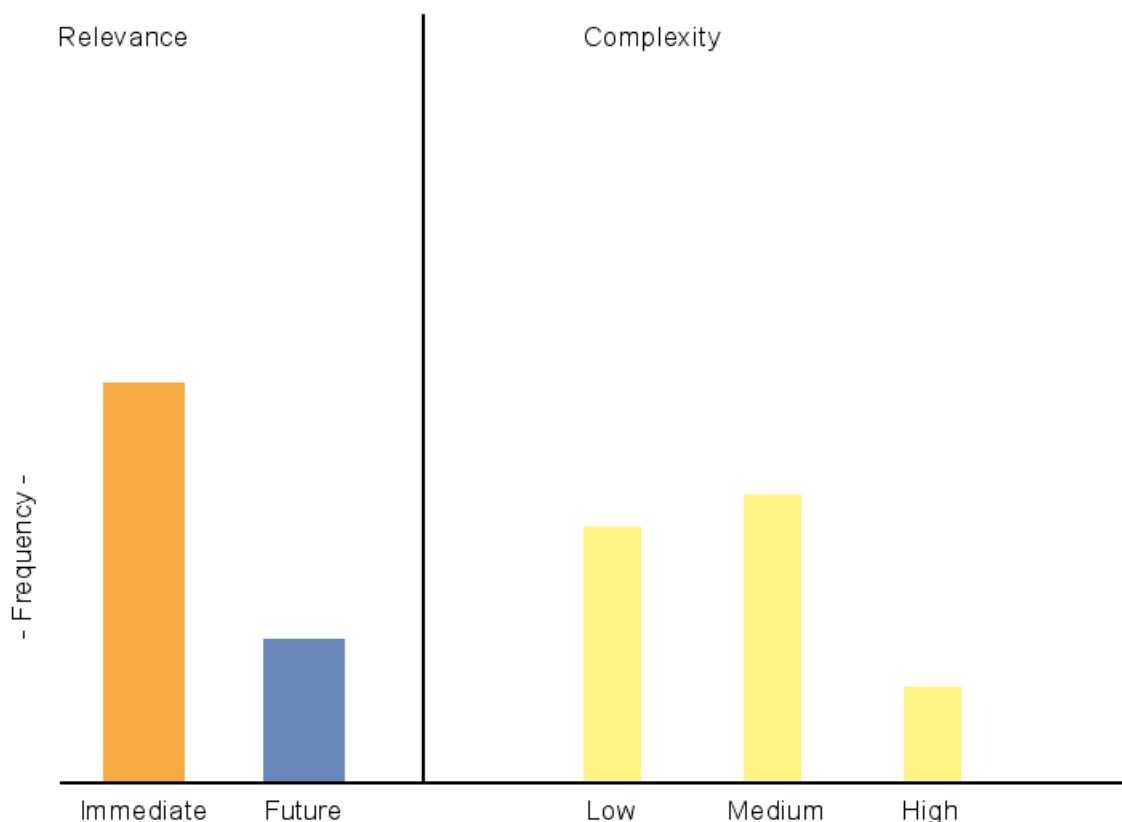
*“unsure whether it is feasible to define or apply mandates in a generic way that allows automation of monitoring. (other than just checking license metadata)”*

*“Metadata with the identification of the version + hasVersion and isVersionOf for linked data.”*

## Issue: Exposing Versioning Information

Repositories include similar versions of documents. In order to make this transparent methods to identify and expose the corresponding information have to be used.

### Exposing Versioning Information



## Comments

*“this depends on software developer communities such as DSpace and EPrints...”*

*“Versioning is closely related to usage statistics, bibliometrics and ‘potentially’ to research data. If different versions of a paper are due to different research data, then it will become crucial to handle this issue. Versioning can be handled by introducing FRBR model into metadata records, but this means a big implementation effort.”*

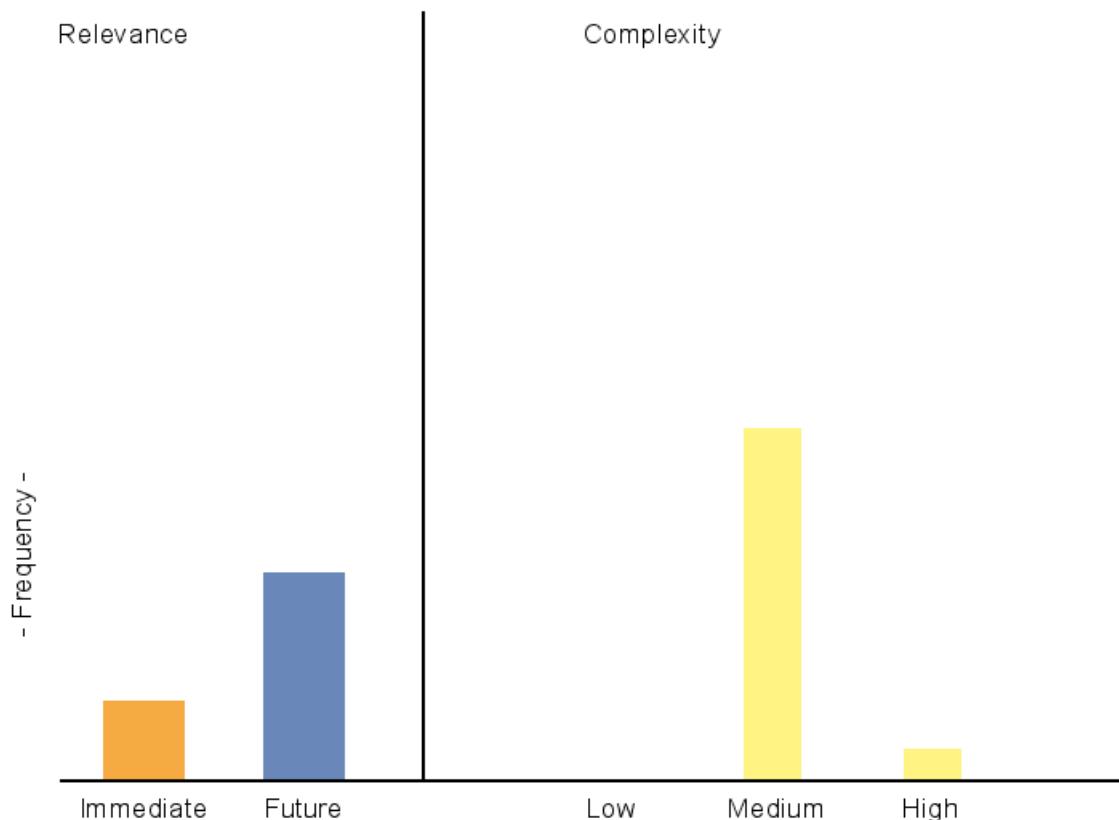
*“Metadata with the identification of the version + hasVersion and isVersionOf for linked data.”*

## Key Aspect: Validation and Aggregation

### Strategic Benefit

Automatic tools for repository managers to optimize their repository infrastructure:

#### Validation and Aggregation



## Comments

*“IMHO, there are not so many repository managers to deal with this...”*

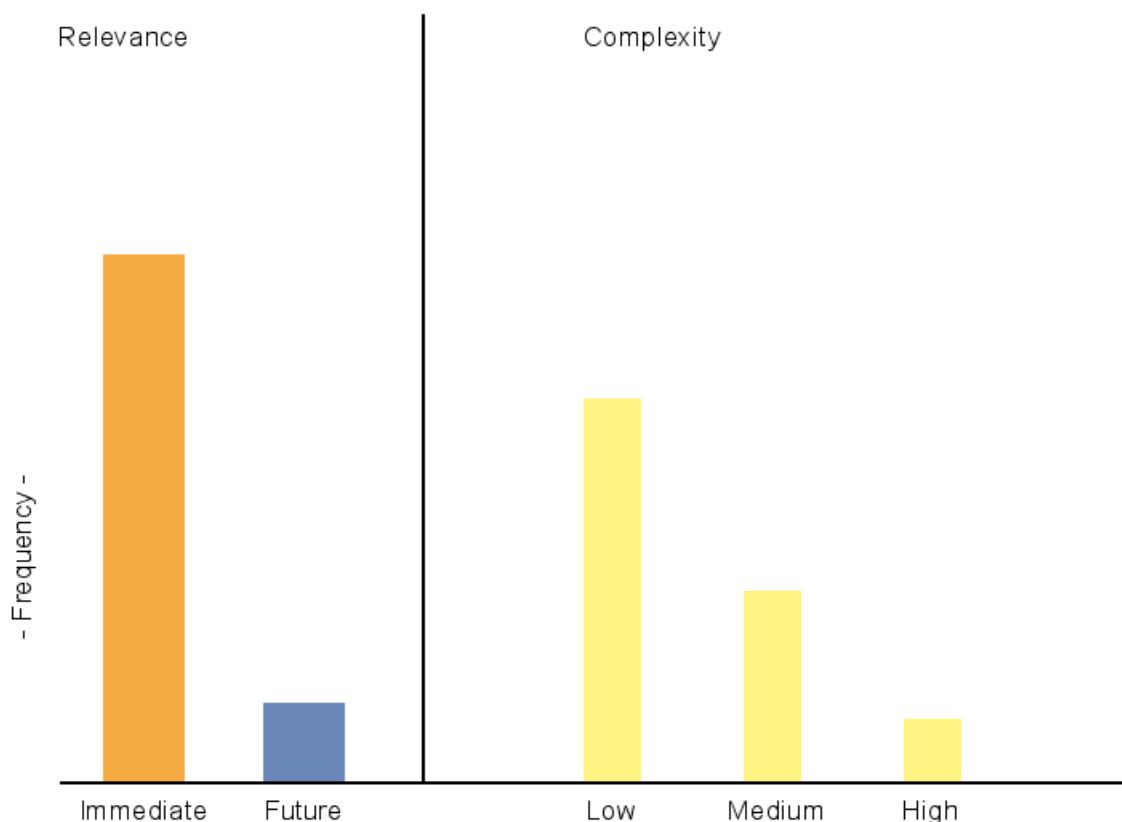
*“Market driven”*

*“There is a need for a new generation of tools that make use of external infrastructure services, exploiting and reusing existing technology. E-Infrastructure projects as EUDAT or agINFRA provide example solutions but this requires changes in the repository software.”*

## Issue: Validating Repository Metadata

Tools for automatic testing of metadata quality (regarding xml schema validation plus guidelines compatibility) can provide valuable information for different purposes in metadata usage and exchange.

### Validating Repository Metadata



## Comments

*“I see validation as a relevant service if there is a particular service against which to validate (i.e. my metadata are valid, compliant for OpenAIRE). As a general validation, I do not see the point. The quality of the metadata is closely related to their capacity to generate good and useful services.”*

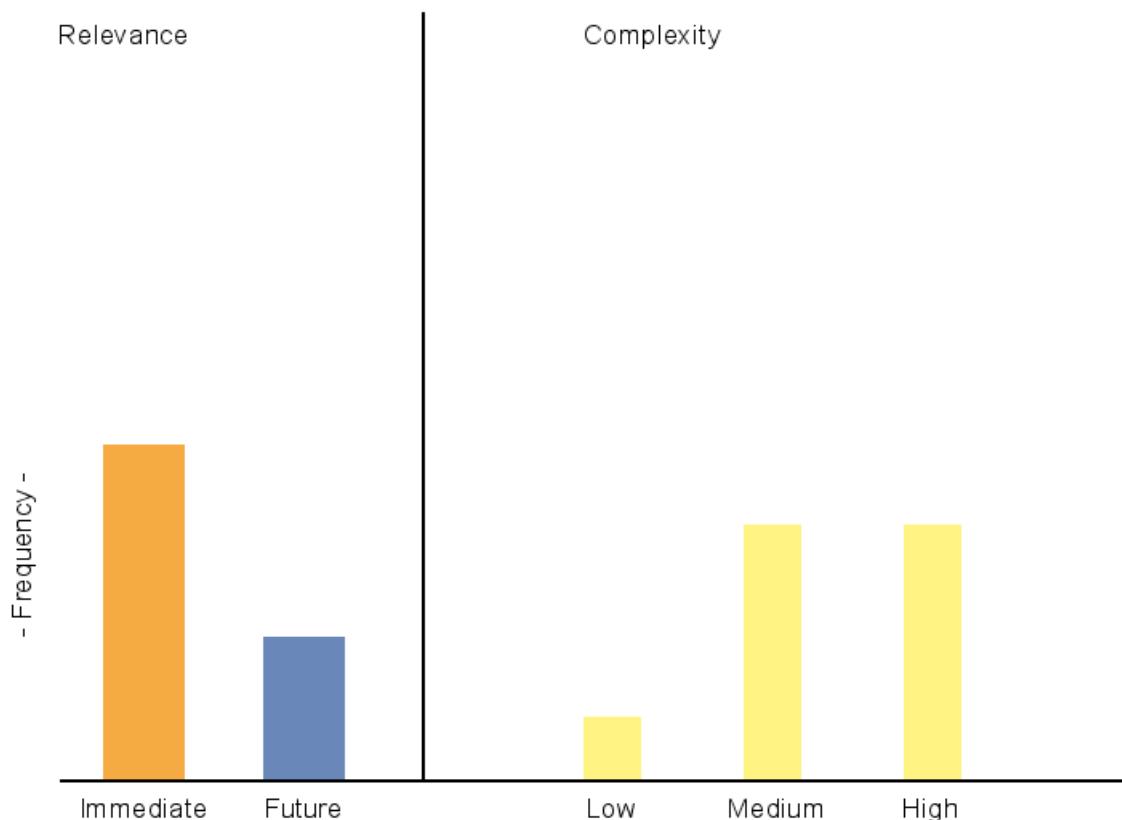
*“Some services are already operational (e.g. DINI).”*

*“At a basic level of completeness and syntactic validation this is easy to implement and would raise the overall quality significantly.”*

## Issue: Processing Related Full Text

Additional repository services can be implemented based on automatic fetching and processing the related full text (allows for example index terms for search purposes, additional metadata extraction, and citation extraction).

### Processing Related Fulltext



## Comments

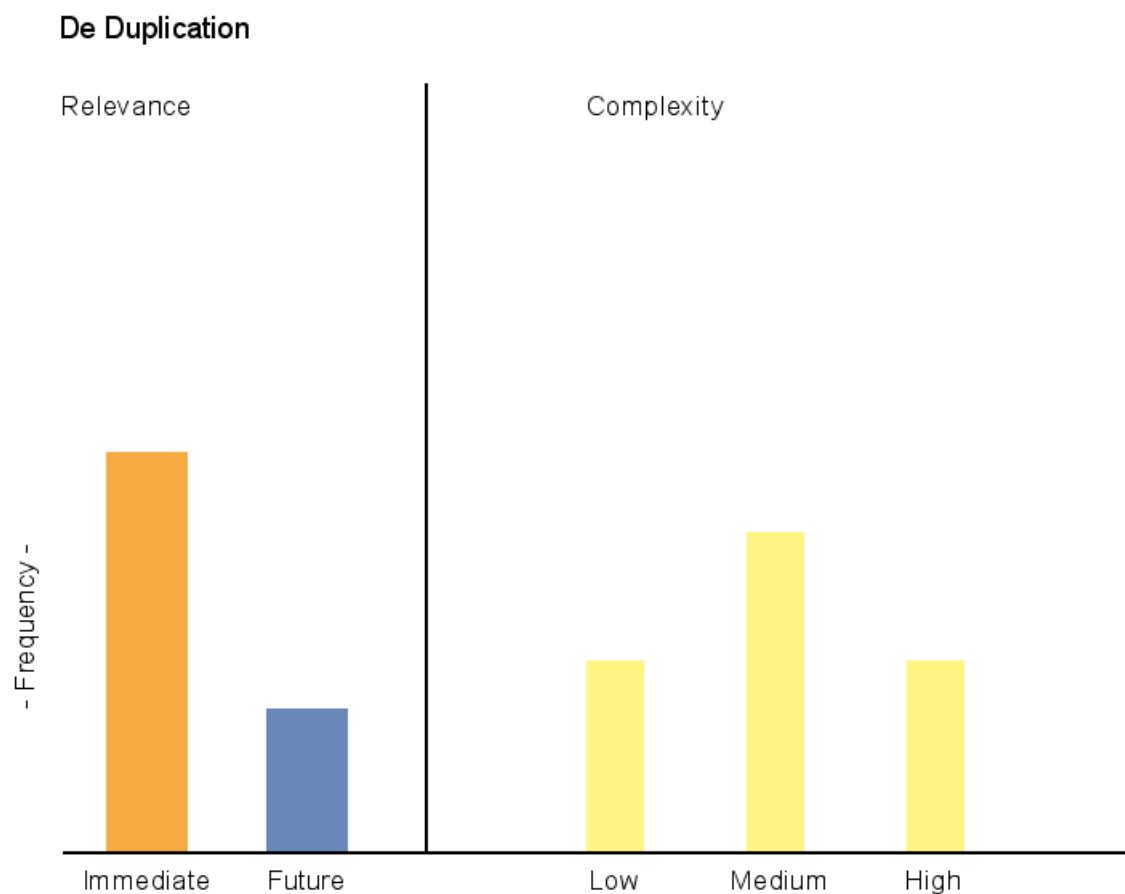
*“Related to guidelines on how to expose full text links and license conditions”*

*“No all-in-one solution, for it implies several single solution (citation extraction and detection, automatic indexing, full text search...).”*

*“This is a big opportunity for automation of metadata enrichment, but the tools would require modifications.”*

## Issue: De Duplication

De-Duplication includes methods to identify and combine identic or similar objects and storing and providing this information through the repository.



## Comments

*“Needs agreed policies and transparency (too many possibilities from a technical perspective)”*

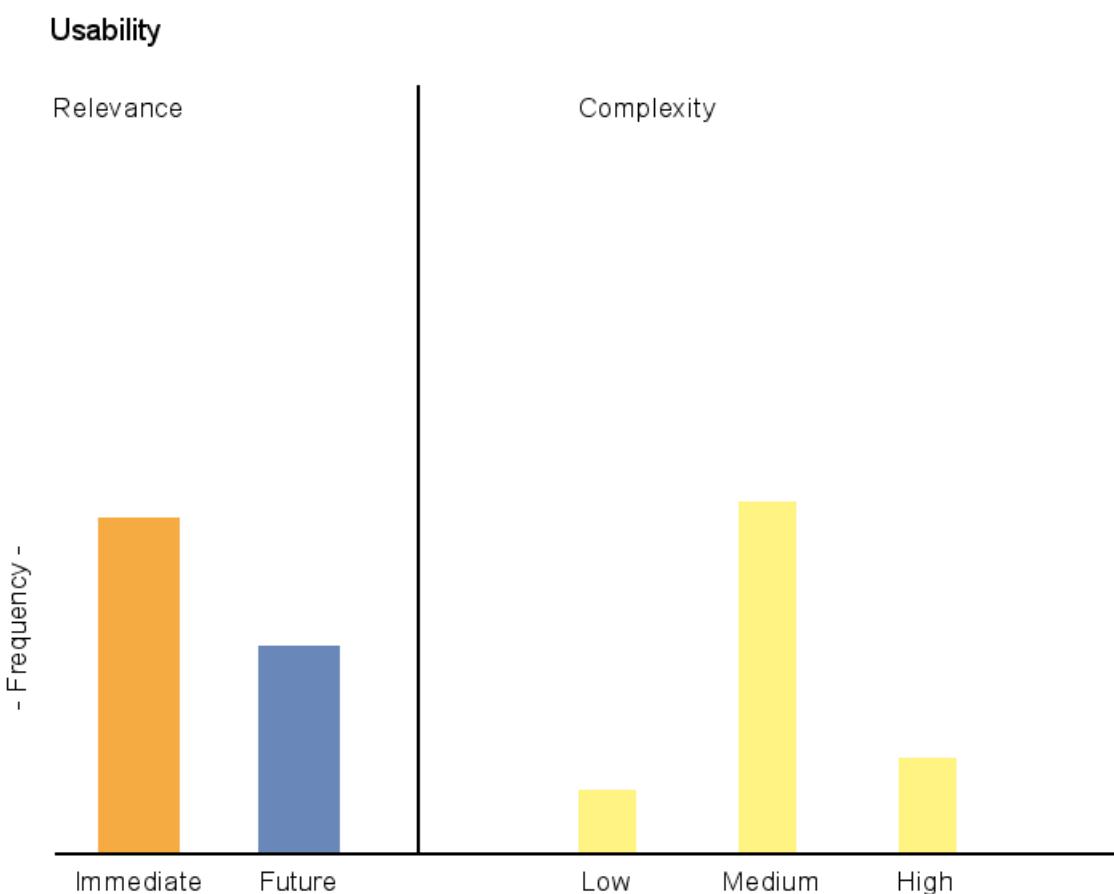
*“Very dependent on quality of metadata, local cataloging rules and almost requires one fixed product line.”*

*“Deduplication technically is not difficult in the part of identifying duplicates. But merging records cannot be done without a consideration of intellectual property. Each metadata record is a piece of IPR, even if it is under open license (moral rights remain). So combination can be done at the user interface, but systems need to preserve the duplicates to acknowledge moral rights and preserve provenance information. This adds management complexity and requires clear policies.”*

## Key Aspect: Usability

### Strategic Benefit

Optimizing end-user usability in order to improve the utilization of repository services:



## Comments

*“Repositories must expose their content to machines in the web; this will be the true usability.”*

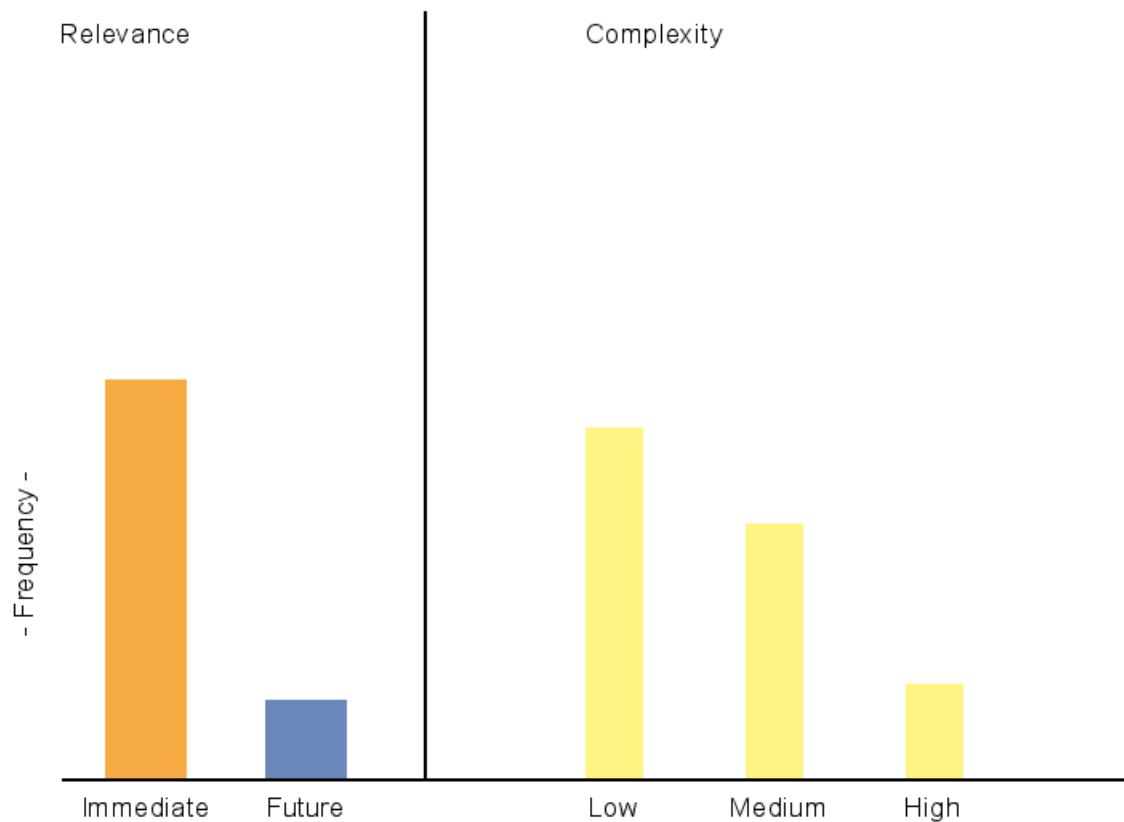
*“Most of a repository’s traffic is initiated from other content aggregators or service providers, e.g. search engines and catalogs. On the other hand, splash pages referred to from external results can become more important to provide more context information (e.g., on similar authors or titles).”*

*“Users need to interact with user services across repositories”*

## Issue: Supporting Authorization and Authentication

Services to handle authorization and authentication (especially in an institutional background) in order to control access and role support.

### Supporting Authorization and Authentication



## Comments

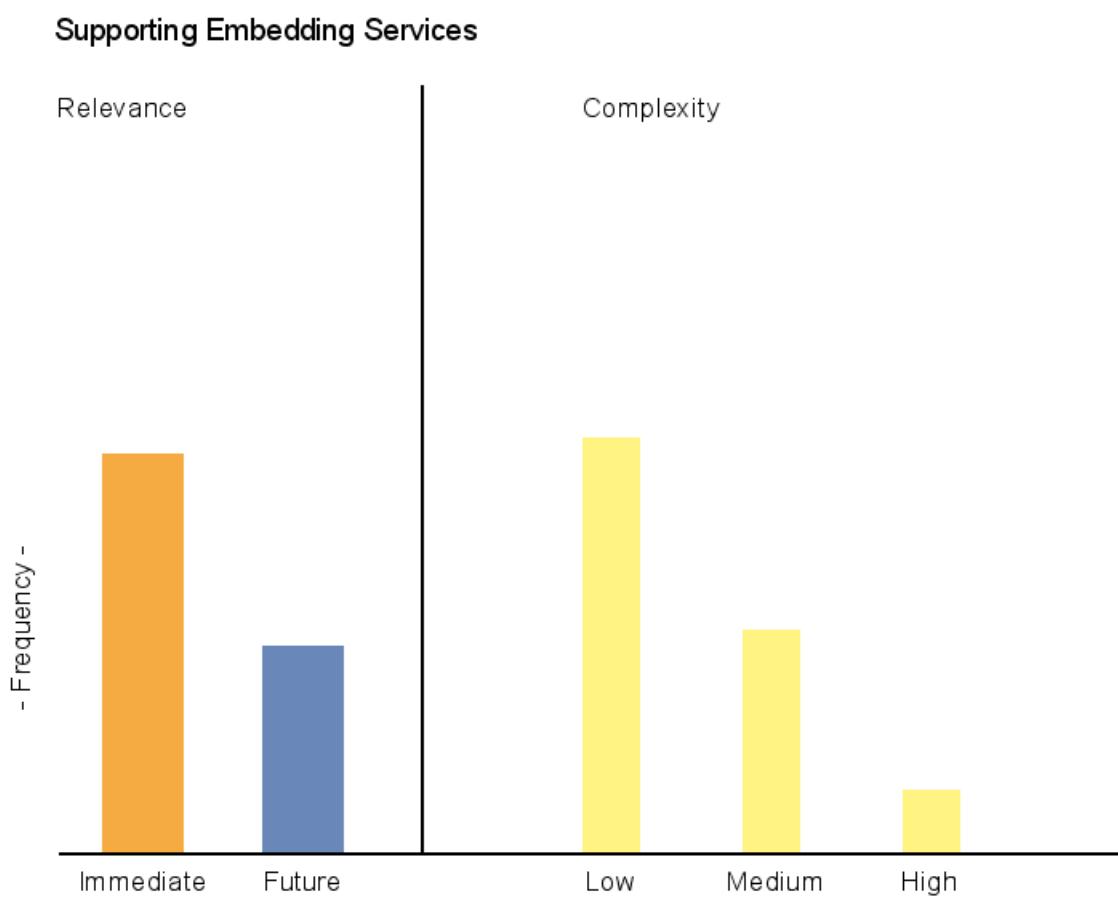
*“Author identification would solve many problems in the development of an OAR.”*

*“Only relevant for institutional repositories with self-upload from researchers. The integration with other campus systems might be a future issue, but I still do not see an absolute mandate for managing one’s scientific output in a repository.”*

*“Stick with national implementations. Implementing international federation is not the task for repositories”*

## Issue: Supporting Embedding Services

Embedding services allow end-users to integrate repository data and functionality in their personal environment. Examples are embedding of publication lists into personal homepages or including search capabilities in institutional web pages.



## Comments

*“Especially when long-term preservation is involved”*

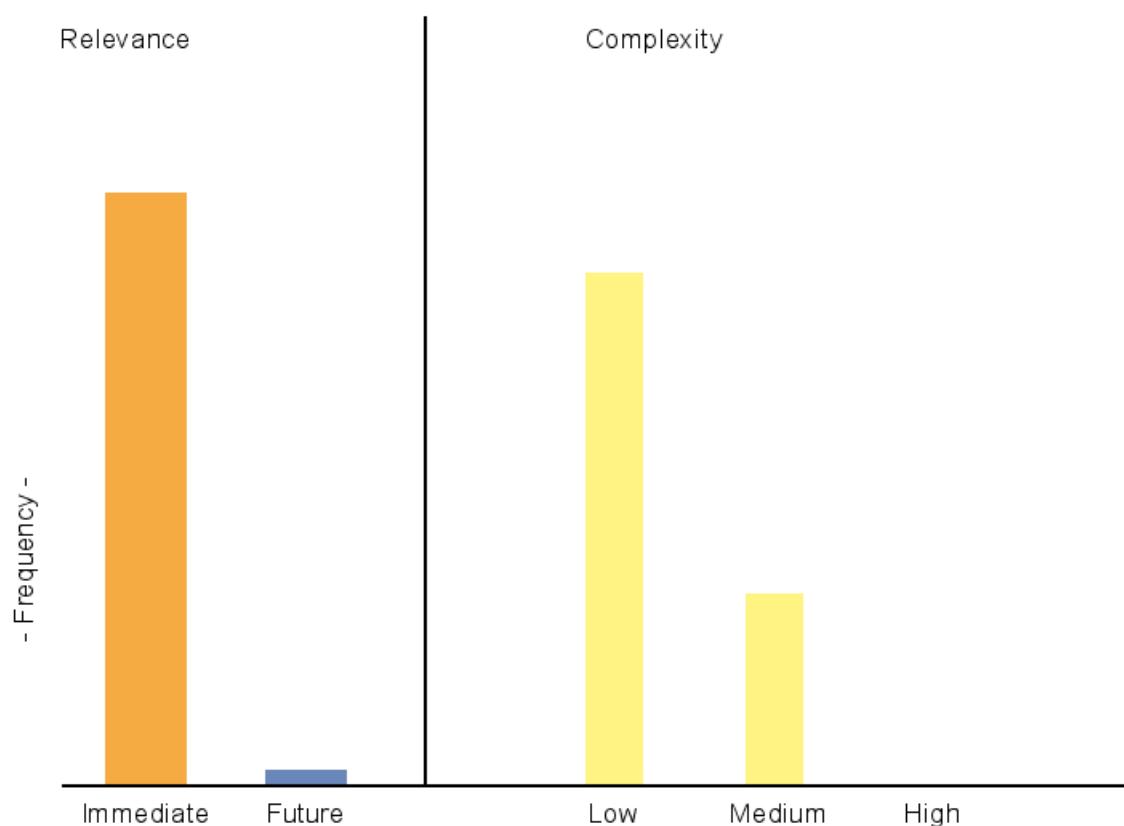
*“Depends on community and scope”*

*“I do not consider this extremely important now, as it required a redeployment of all the current infrastructure in LOD, and this will take time.”*

## Issue: Exposing Publication Lists

Personal and institutional publication lists can offer a service of institutional repositories based on the stored bibliographic metadata and relations with persons and institutions.

### Exposing Publication Lists



## Comments

*“Normally, these publications only contain a subset of all publications (from persons or institutions), hence it will be a crucial question how to enrich and integrate these lists with complementary data.”*

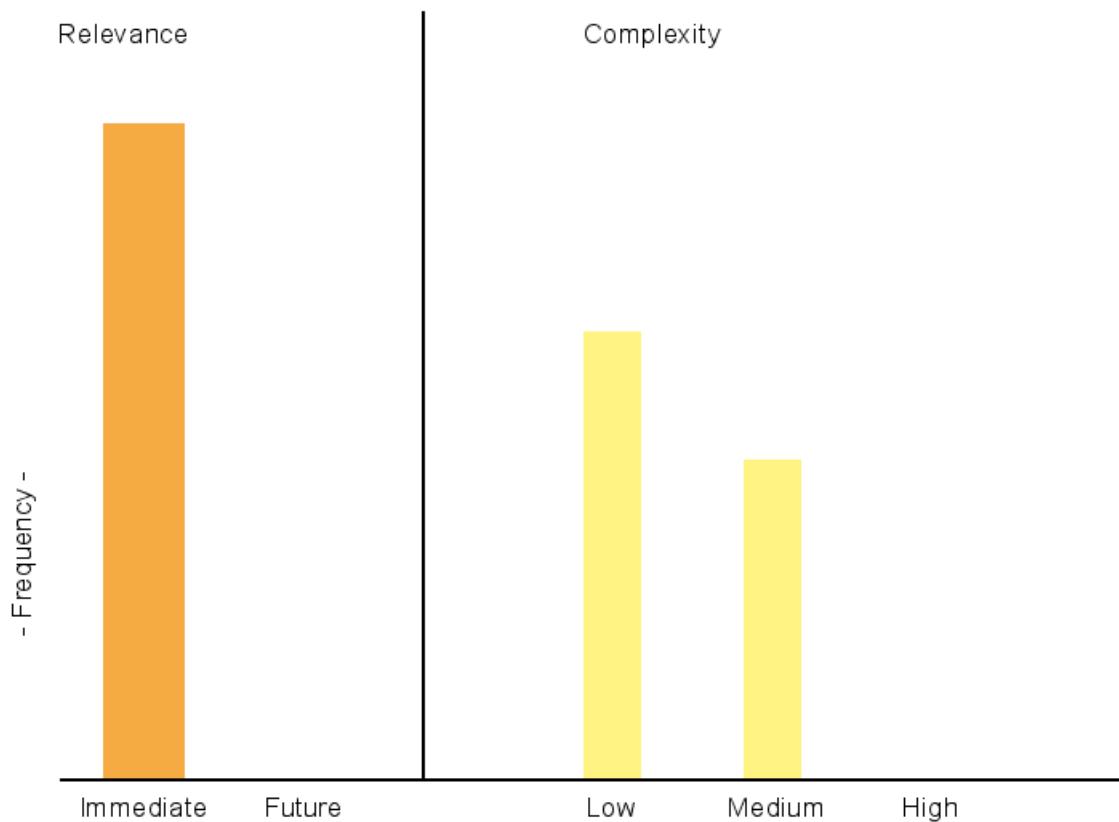
*“Sustainability depends on local management of authority records for persons, organizational units etc.”*

*“Very important”*

## Issue: Exposing Citation Formats

Bibliographic metadata are cited in various formats. In this context repositories may expose their metadata in those citation formats. There are technical implementations available to integrate the creation of numerous popular citation formats.

### Exposing Citation Formats



## Comments

*“Should be standard since it is one of the few real services that a repository could possibly offer”*

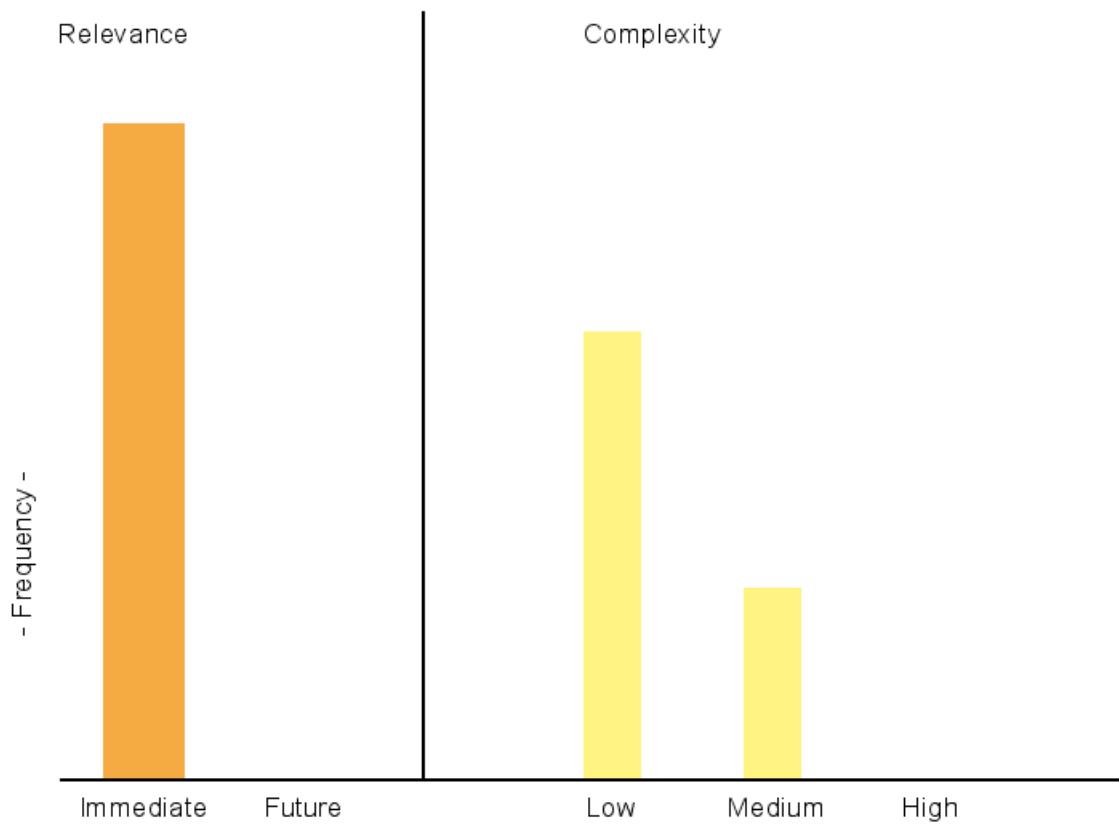
*“Personally, I do not know anybody using these citation strings for reference management... Normally, repositories are not conceived as environments for reference management, rather catalogs or portals.”*

*“Only applies to research items”*

## Issue: Supporting Data Export functions

Export functionality delivers the metadata in order to process them elsewhere later and thus contains a broader range of metadata information than citations only.

### Supporting Data Export functions



## Comments

“Should be standard since it is one of the few real services that a repository could possibly offer”

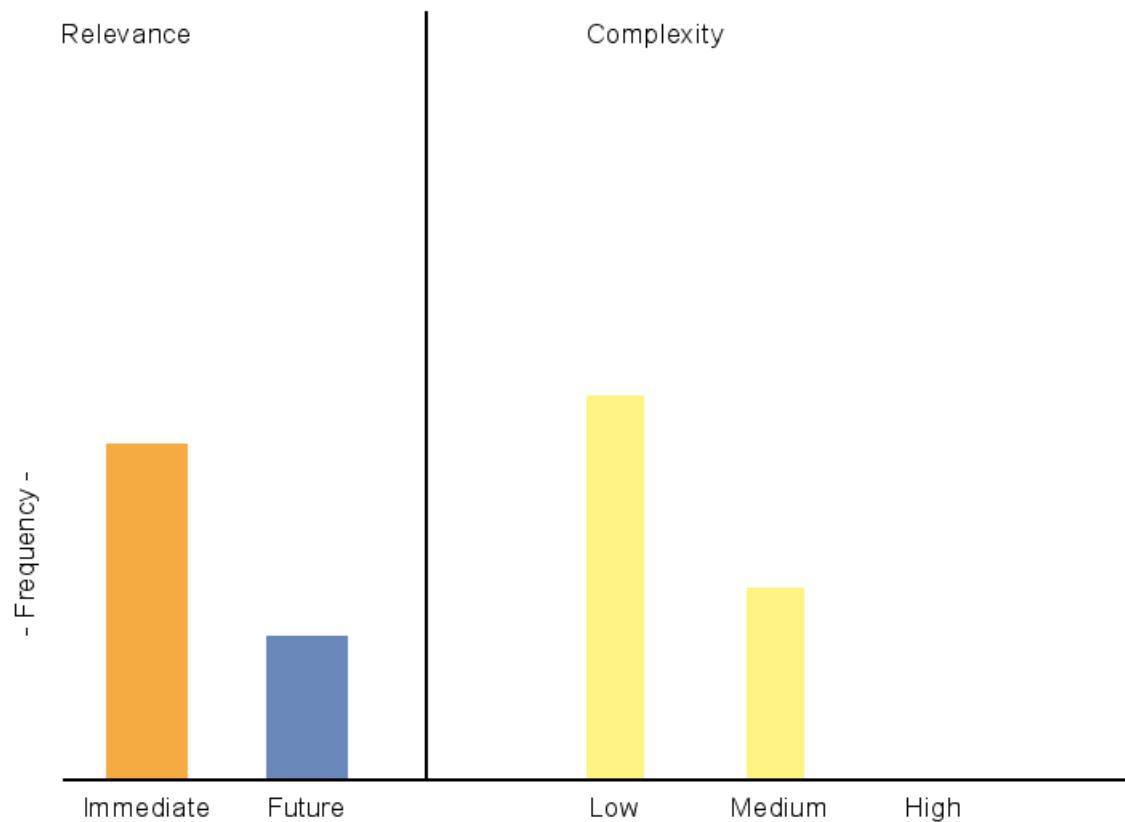
“Depends on community and scope”

“These functionalities are already implemented in most bibliographic systems.”

## Issue: Integrating Availability Services

Link Resolving services compute information about the availability of objects, thus this service can be relevant for repositories in a bi-directional approach e.g. adding availability links and playing a role as target system for storing resources.

### Integrating Availability Services



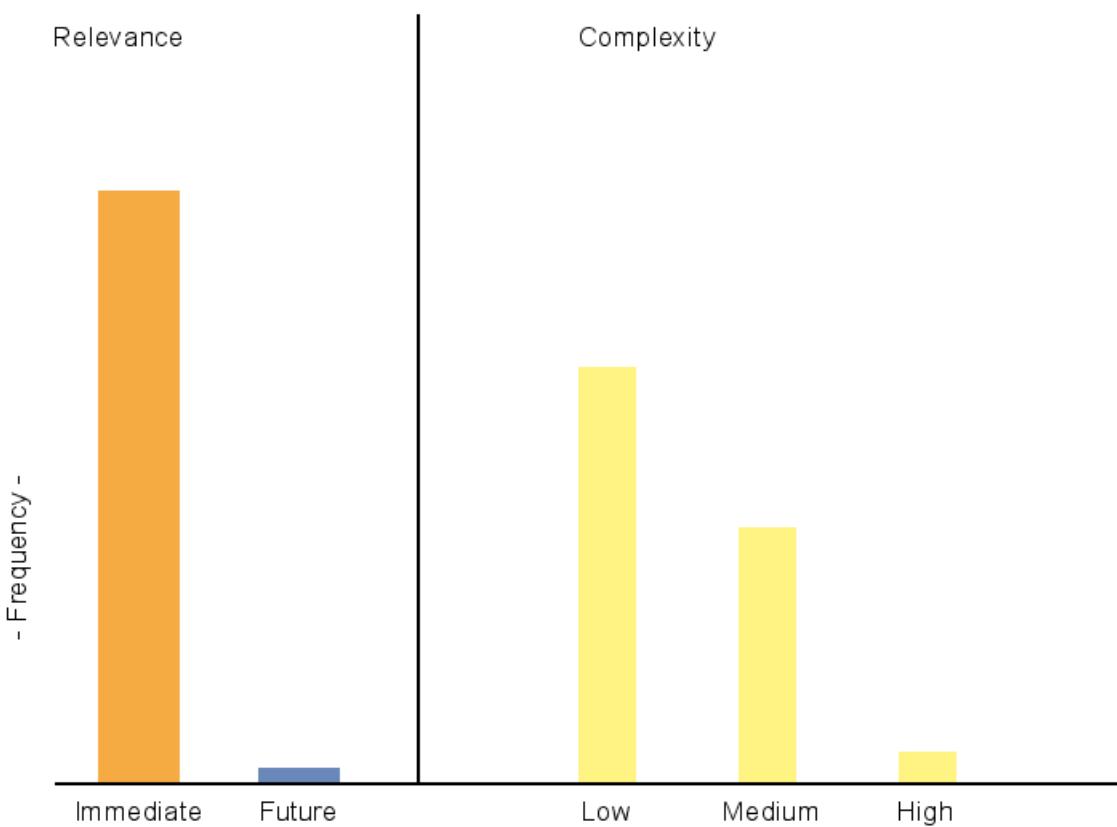
## Comments

*“More exposure for the commercial competition?”*

### Issue: Supporting Author Identification Systems

Author identification systems intend to identify academic authors in a unique way. Different systems are available (ORCID, ResearcherID, AuthorClaim etc.) and can be used in the repository context.

#### Supporting Author Identification Systems



## Comments

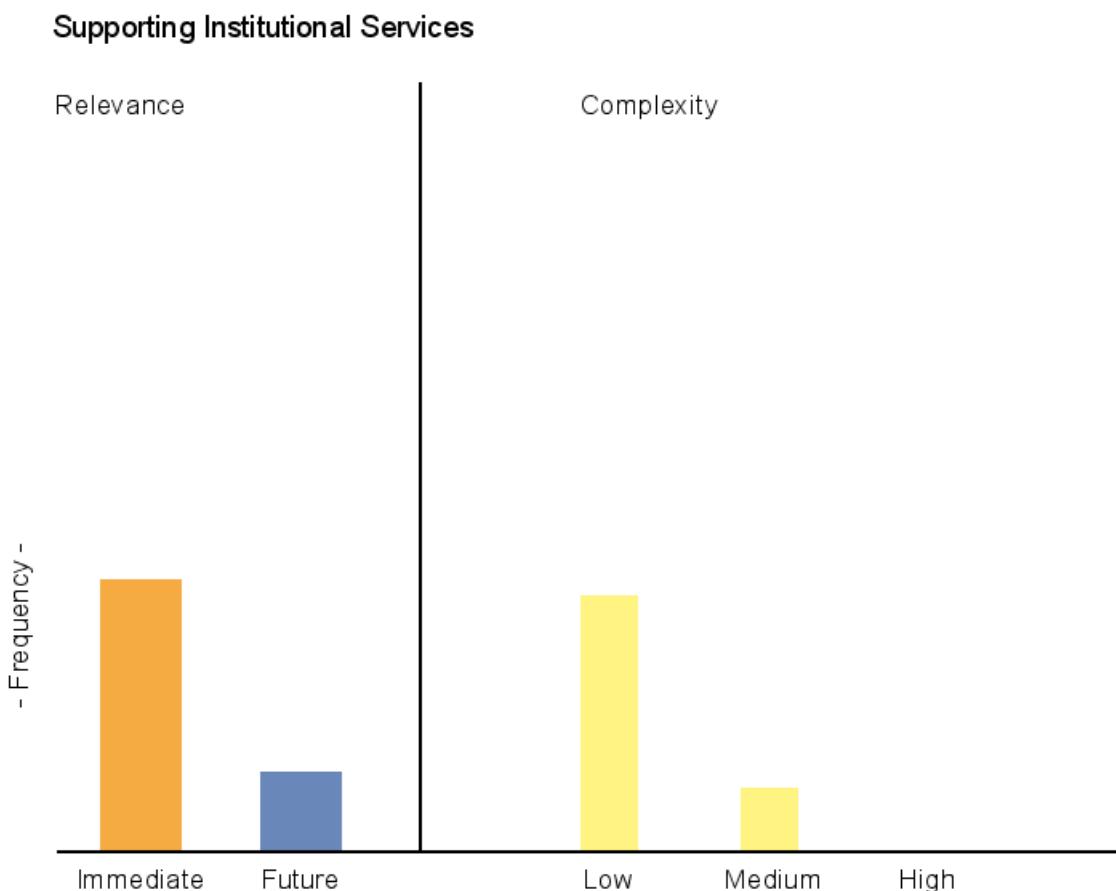
*“IMHO, a very important extension of current repositories, but relying on external authority data maintained by external systems / staff; needs a lookup service for persons, which can be applied during indexing and searching of data”*

*“Clearly supports open systems like ORCID instead of proprietary ones”*

*“Trivial to do with BASE and AuthorClaim”*

### Issue: Supporting Institutional Services

Repositories contain institution-related and administrative information. Specific services should be able to provide and to display this.



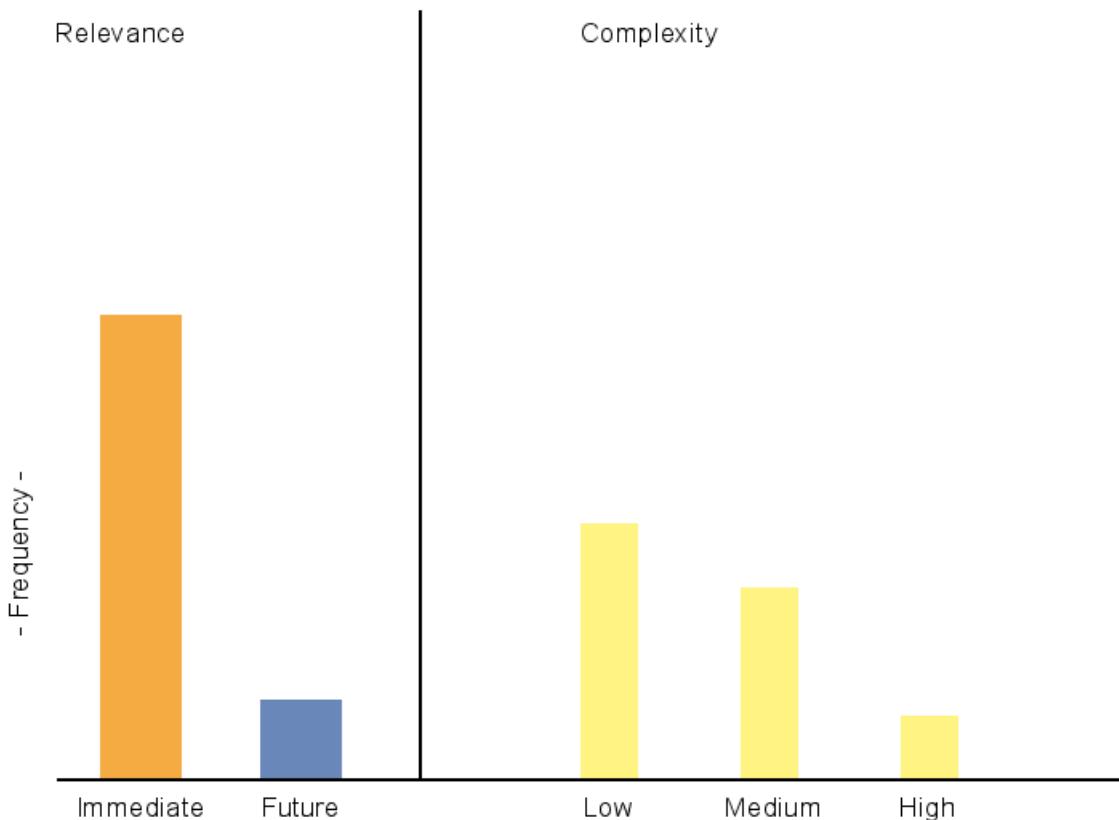
### Comments

"Depends on scope"

### Issue: Extending End-User Usability

This issue includes all end-user orientated functionality and services (search, open-access transparency, drilldown etc.) in order to upgrade the repositories' usability.

#### Extending End-User Usability



#### Comments

“Faceted search would be one of those functionalities, but there are not too many of them.”

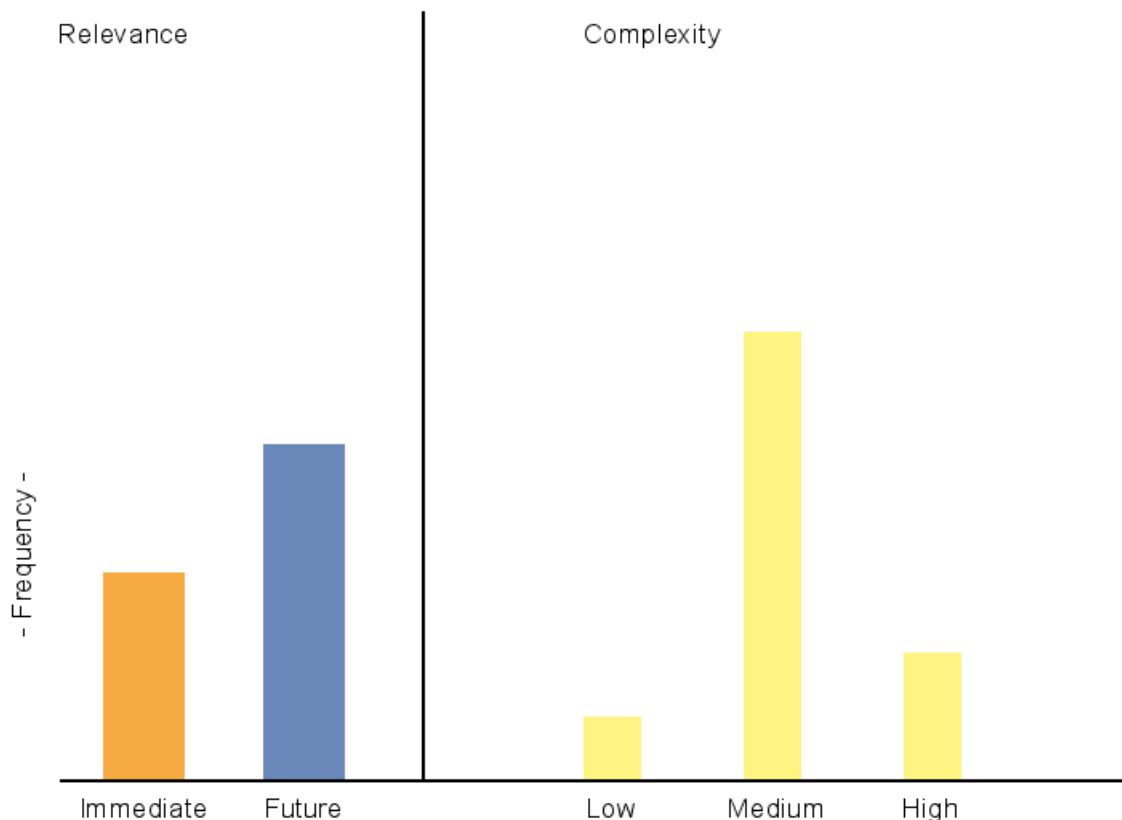
“Important, but an interoperability-issue? Are best-practices sufficient?”

"We need cross repository services"

### Issue: Extending Usage of Visualization Tools

Repositories contain complex structures of metadata, contents and additional information and especially interconnections between them. Visualization tools may help to make this information more transparent and understandable and accordingly optimize end-user interoperability.

#### Extending Usage of Visualization Tools



#### Comments

*“But I do not think visualization is a repository issue; our issue is to expose content in a way that can be integrated (from distributed sources) used, visualized, etc. in a seamless way in any end user environment.”*

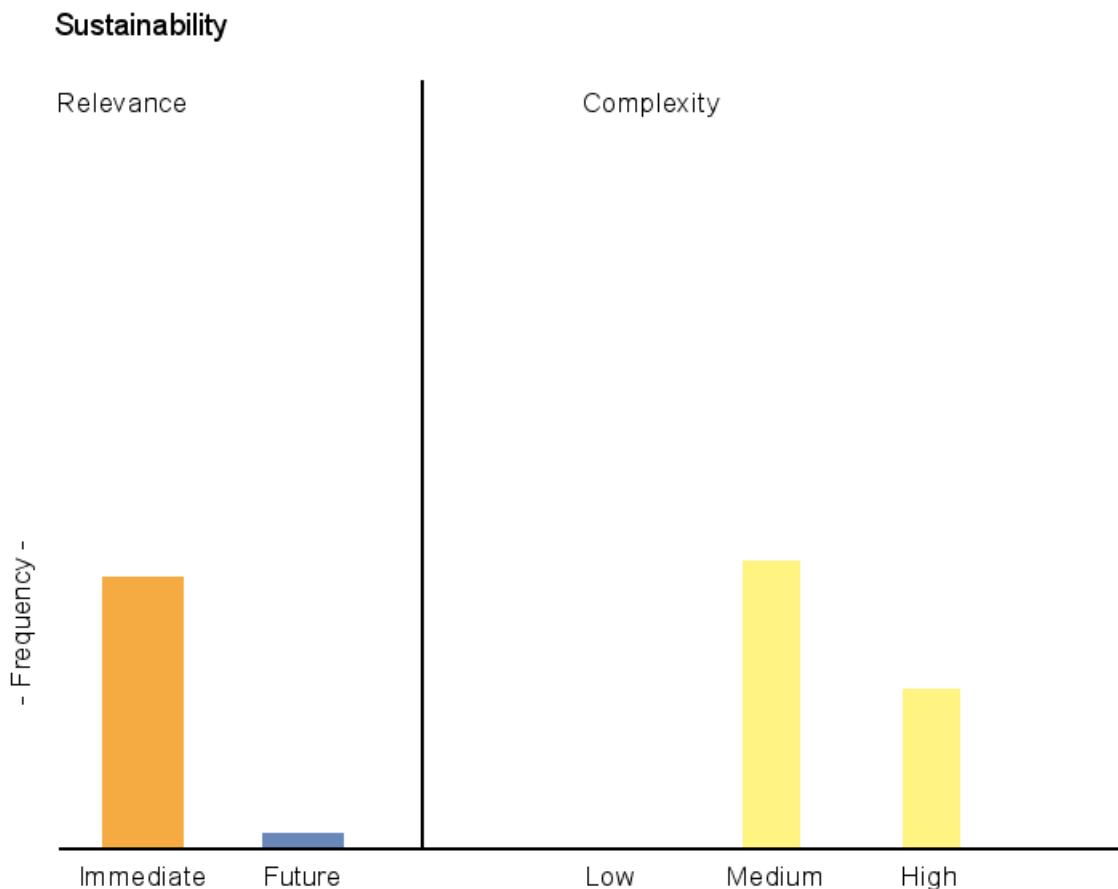
*“Needs good data quality in the first place”*

*“Depending on the functionalities and size; cross-repositories interactive visualization tools should be interesting as well because they can expose relations between researchers/institutions”*

## **Key Aspect: Sustainability**

### **Strategic Benefit:**

Supporting the Need for stable and Persistent Access of Repository Services

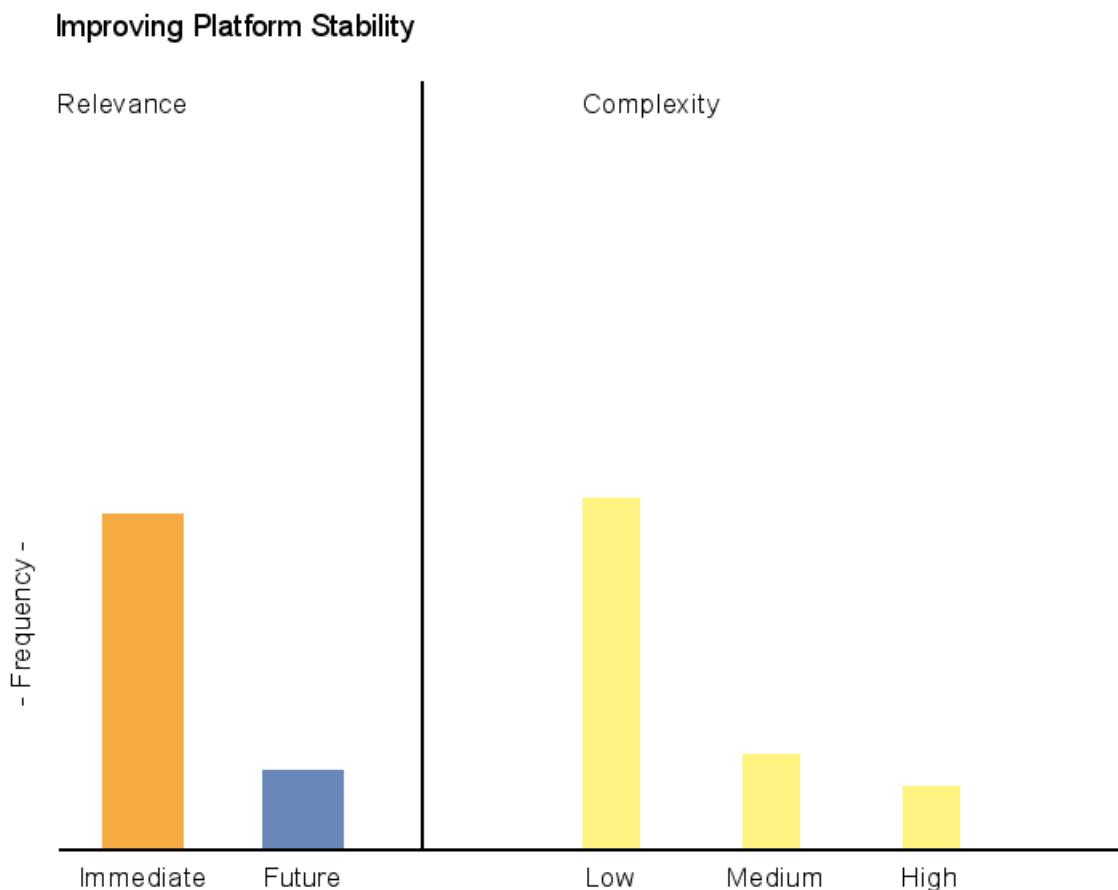


### **Comments**

*"This is an issue of economic models with no universal or clear solution."*

## Issue: Improving Platform Stability

Technical activities to supervise and maintain platform availability (including core system, interfaces, APIs):



## Comments

*“Can be overview of best practices / guidelines”*

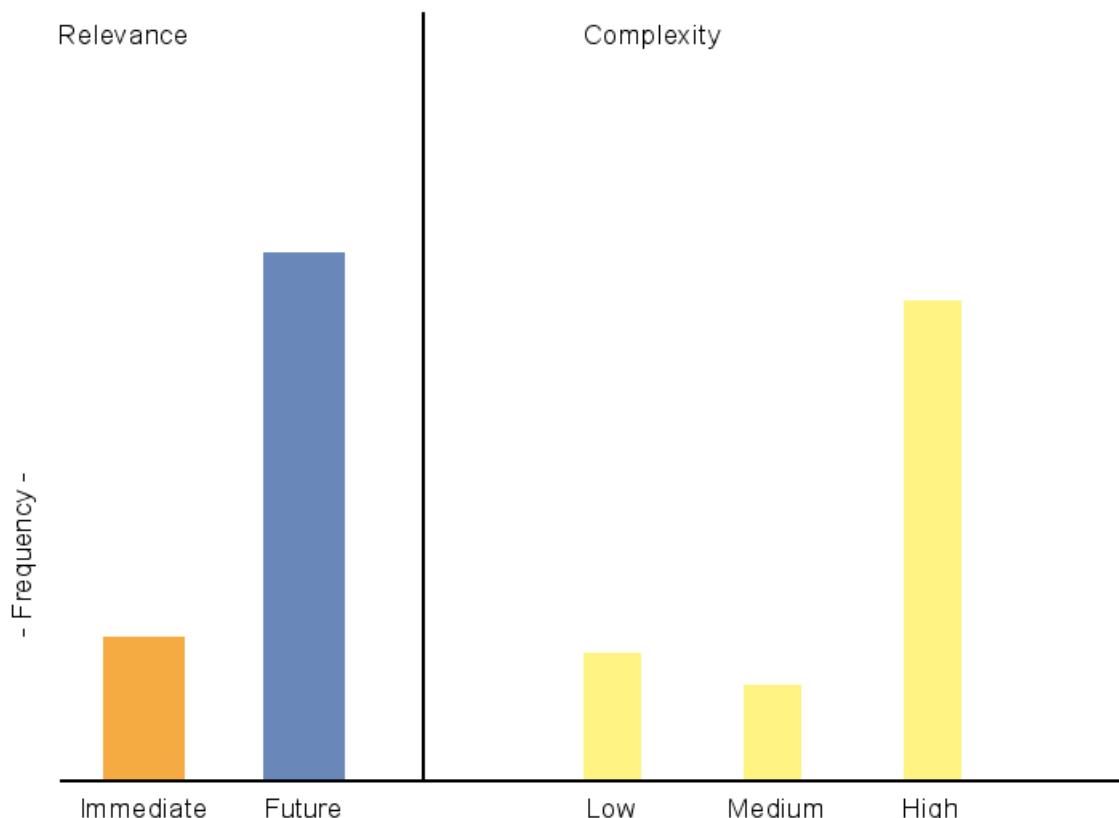
*“Has to be part of daily operations”*

*“I don’t consider this an issue for the future, but some basic management daily work.”*

## Issue: Supporting Long-term Preservation and Archiving

Long-term preservation includes efforts to ensure continued access to digital materials (regardless of technology changes over time while archiving includes access services to the material).

### Supporting Long-term Preservation and Archiving



### Comments

*"LTP is more a theoretical than a practical requirement, but nonetheless an important one. As for research data, a solution for LTP should be considered outside of the repository, in conjunction with other data sources and media formats."*

*"Difficult to implement by a repository itself, but potentially easy to outsource to dedicated LTP organizations"*

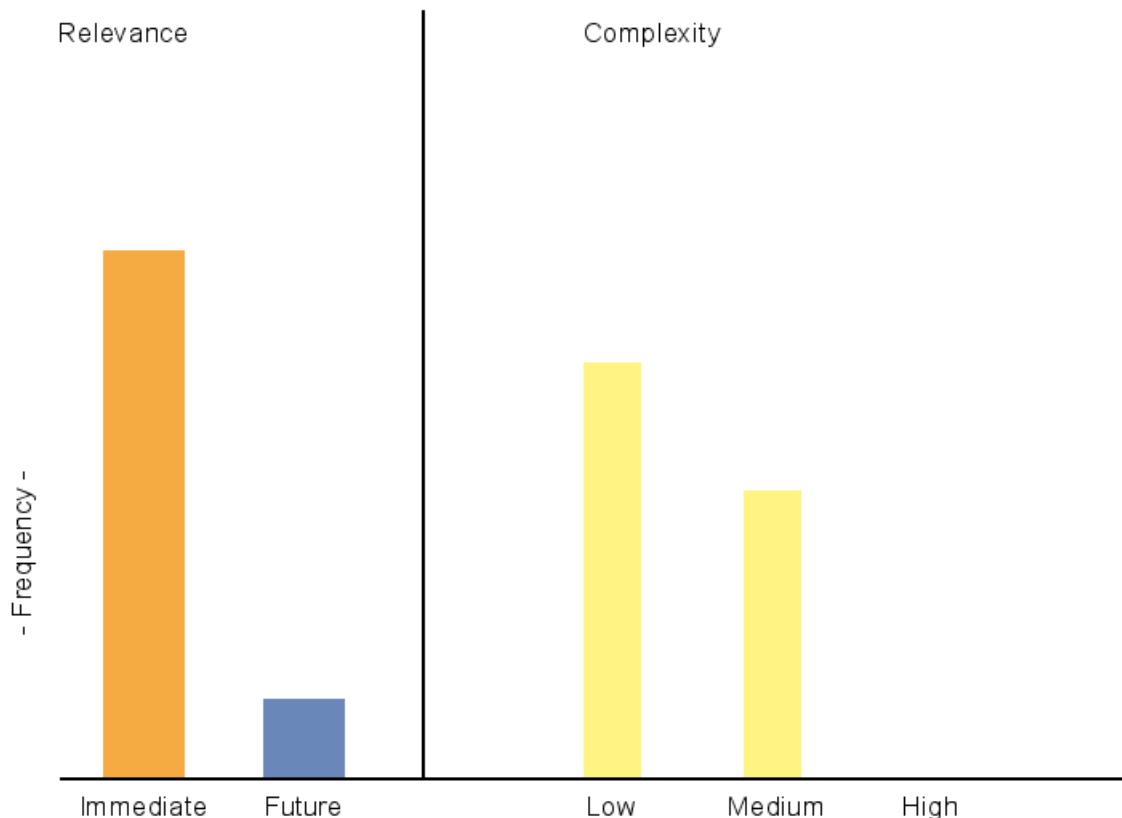
*"LTP is more a theoretical than a practical requirement, but nonetheless an important one. As for research data, a solution for LTP should be considered outside of the repository, in conjunction with other data sources and media formats."*

*"Keep things running."*

## Issue: Exposing Persistent Identifiers

Persistent identifiers for digital objects are unique stable references to hide eventual local changes. There are different systems in use to implement the needed resolving service.

### Exposing Persistent Identifiers



## Comments

*“Any kind of additional, persistent identifier is useful to fetch more, contextual? Information, e.g. PMID/PMCID, ISSN, DOI”*

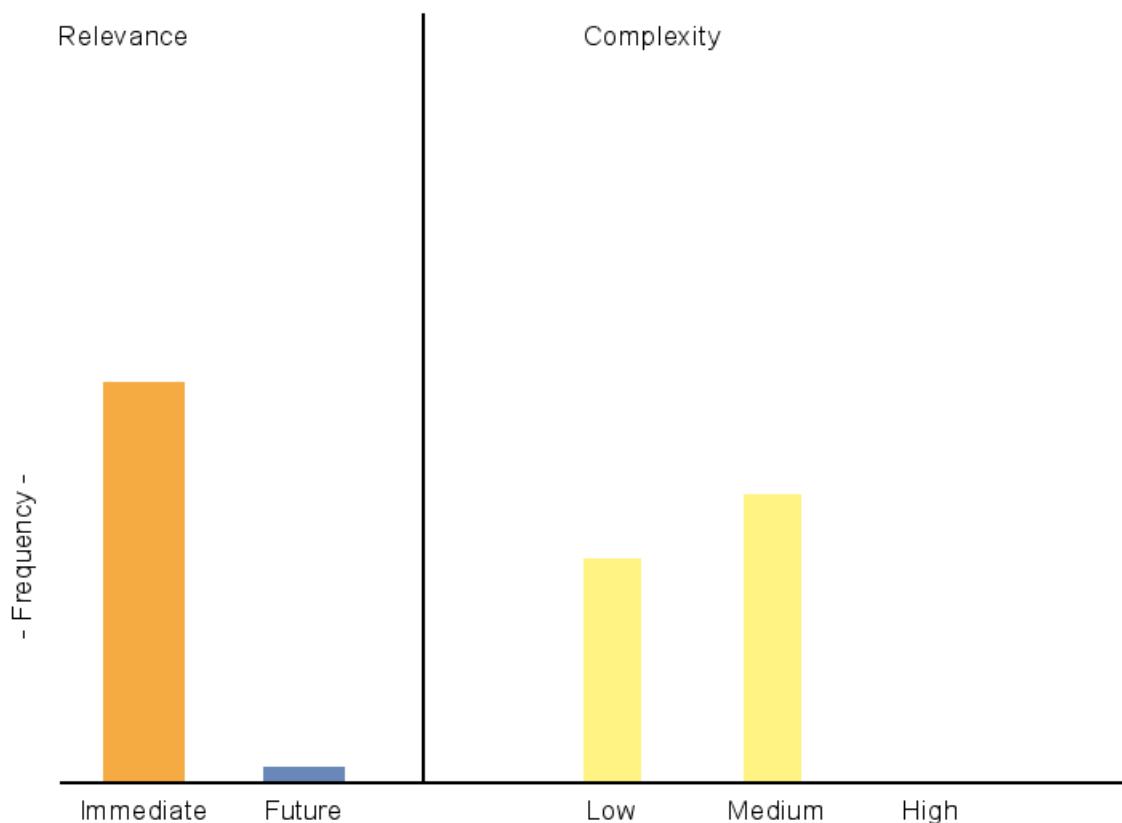
*“Agree, there is a need of an international resolving service, but this issue is beyond repositories”*

*“Assigning PID is easy, but ensuring the ‘promised’ persistence is challenging”*

### Issue: Integrating different Persistent Identifiers

Repositories do often provide their own identifier system, which is applied to the objects which are generated through this repository. With the increasing exchange (import/export/harvesting/LOD) of metadata, different identifier systems will ‘clash’ in a repository environment.

#### Integrating different Persistent Identifiers



### Comments

*“I do not know if this is really an issue for repositories, but for data aggregators and service providers. For repositories, there might be the solution with versioning.”*

*“Very relevant issue for us (URN and DOI)”*

*“Need to support interoperability of different PIs”*

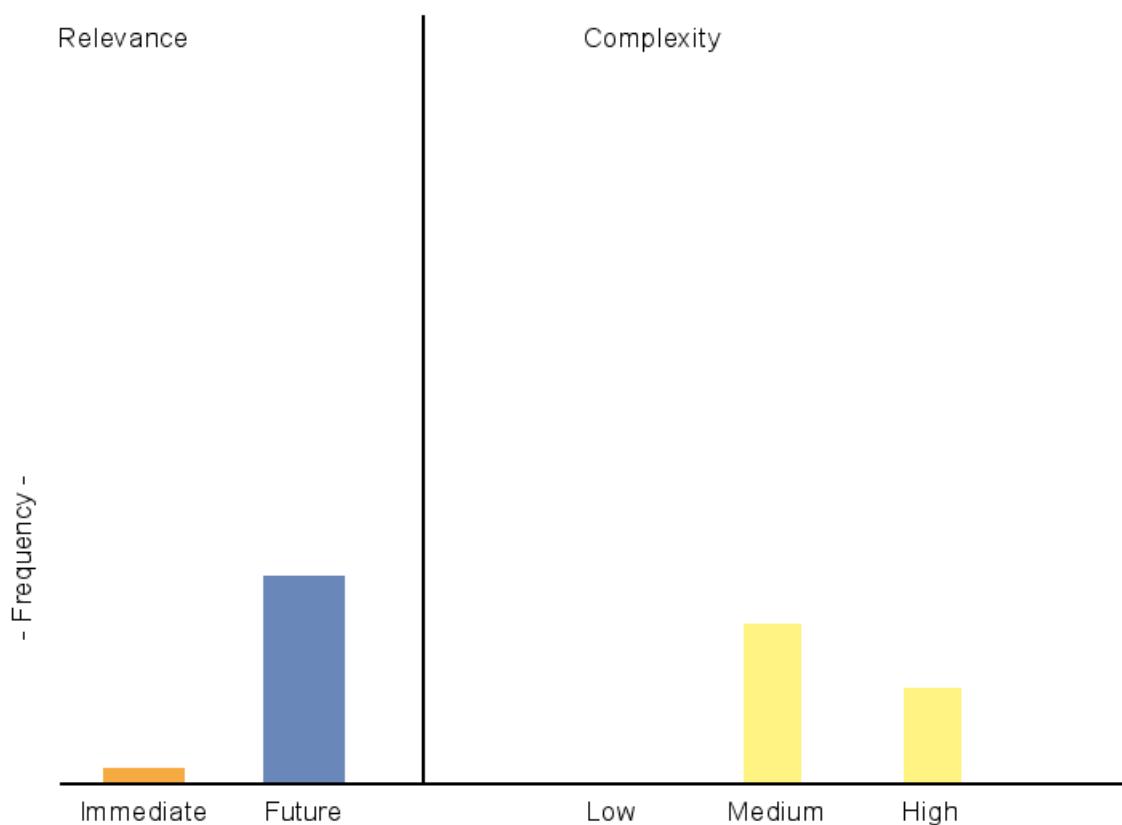
## Key Aspect: Technical Issues

### Strategic Benefit

Extending the Bi-directional Connectivity of Repositories and their Services

Repository systems act more and more in a multi-faceted environment with different other systems (especially CRIS, VRE, Publication Management, Publishing Platforms and several others). This approach implicates the strong need for connectivity efforts to implement the communication channels.

#### Technical Issues



### Comments

*“Important future aspect of an OAR”*

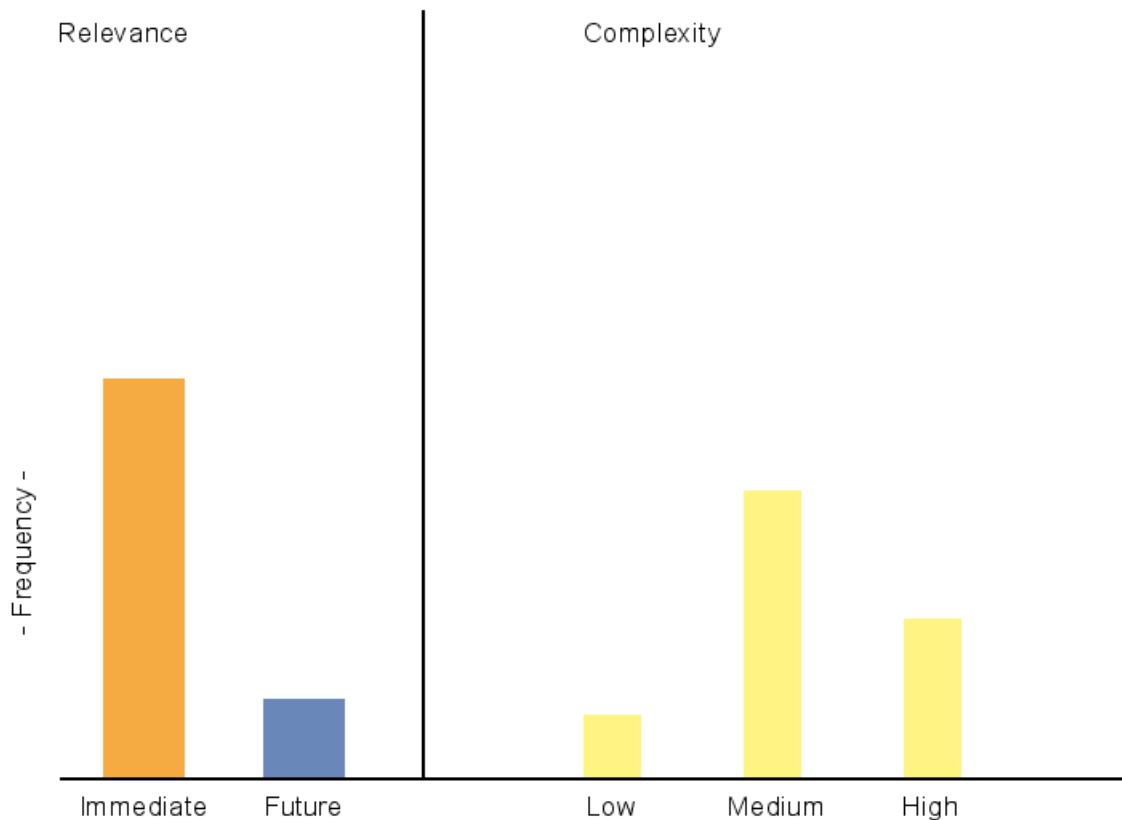
*“Repositories as web service architecture; exposing content as linked data”*

*“Central issue for development of scholarly communication ecosystem”*

## Issue: Defining Architectural Recommendations for Repositories and their Interoperability

In this context architectural, technical and strategic settings in order to establish for bi-directional connectivity in this environment are of fundamental relevance.

### Defining Architectural Recommendations for Repositories and their Interoperability



## Comments

“Web services”

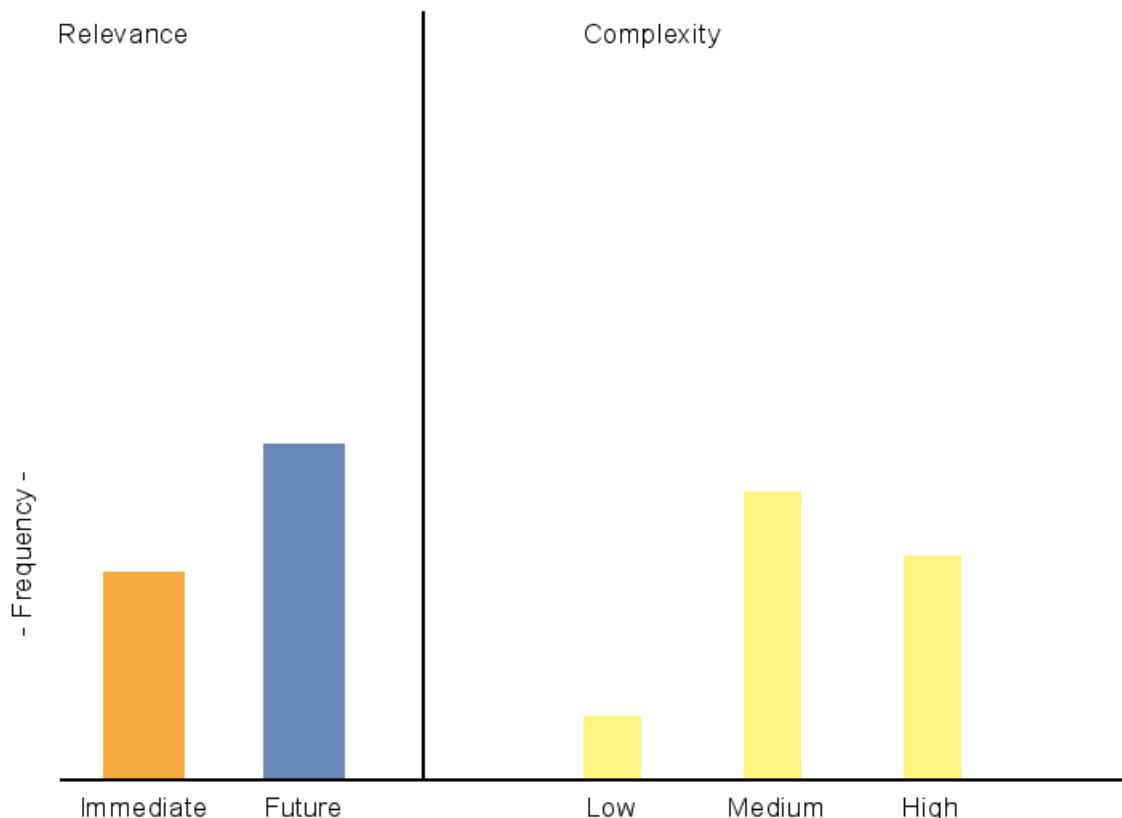
“Of high importance”

“This is very important and pressing. There is a need to advance beyond OAI-PMH and DC to richer formats and APIs not relying on harvesting.”

### Issue: Extending/Replacing Metadata Exposition Protocols

Currently OAI-PMH is the standard for metadata exchange. Since longer there is a discussion about extending the protocol or using other protocols for the purpose of improving the data exchange (OAI-PMH 3.0, ResourceSync for example).

#### Extending/Replacing Metadata Exposition Protocols



### Comments

*"Establishing another protocol on top of HTTP is a tricky issue, requiring lots of personal work"*

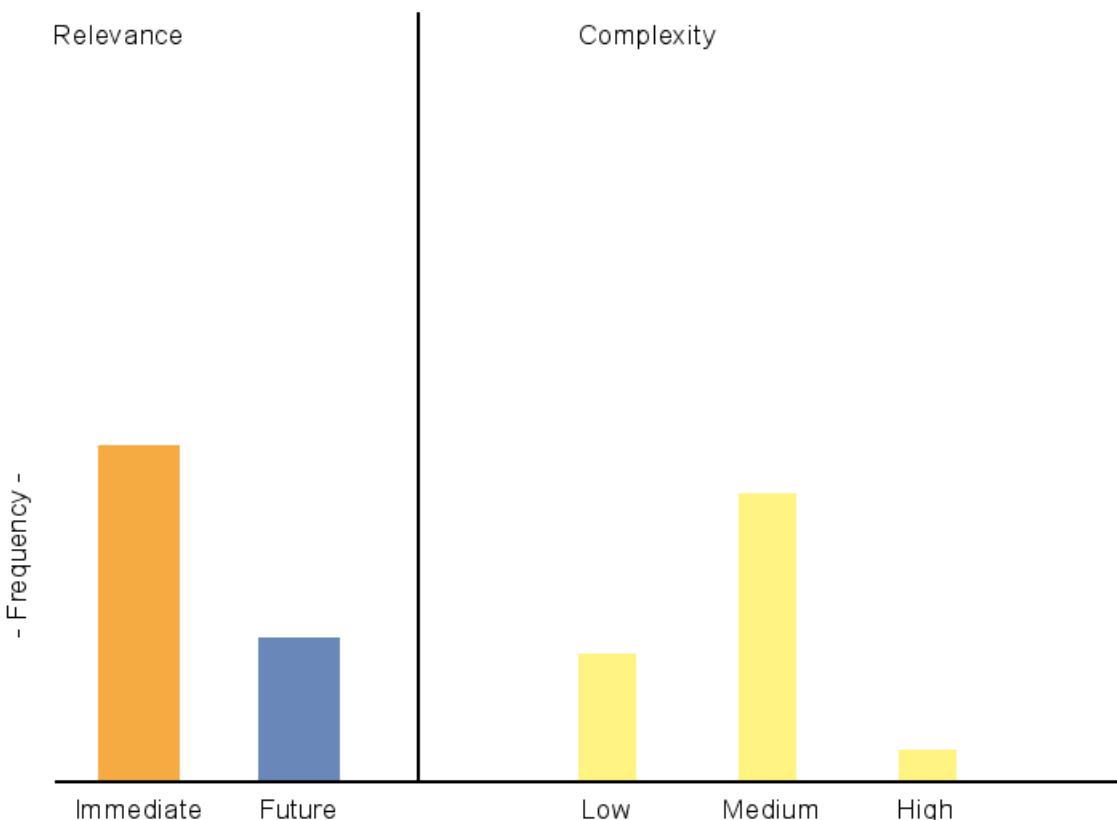
*"OAI PMH is a 'community standard' because of his simplicity and because it can be 'misused' (Like TeX). There should be complimentary protocols for other purposes and scope."*

*"I would say 'complementing'. OAI PMH is good at was it designed for. But there are other needs that could be solved by different approaches, e.g. REST interfaces or registries."*

## Issue: Supporting OAI Service Provider Usage

Repositories as OAI data providers interoperate with OAI Service providers by delivering their metadata and making their characteristics transparent (contact, contents, open access status on repository and/or document level, used classifications, coverage). The coverage and grade of the delivered information has a certain influence on OAI communication quality.

### Supporting OAI Service Provider Usage



## Comments

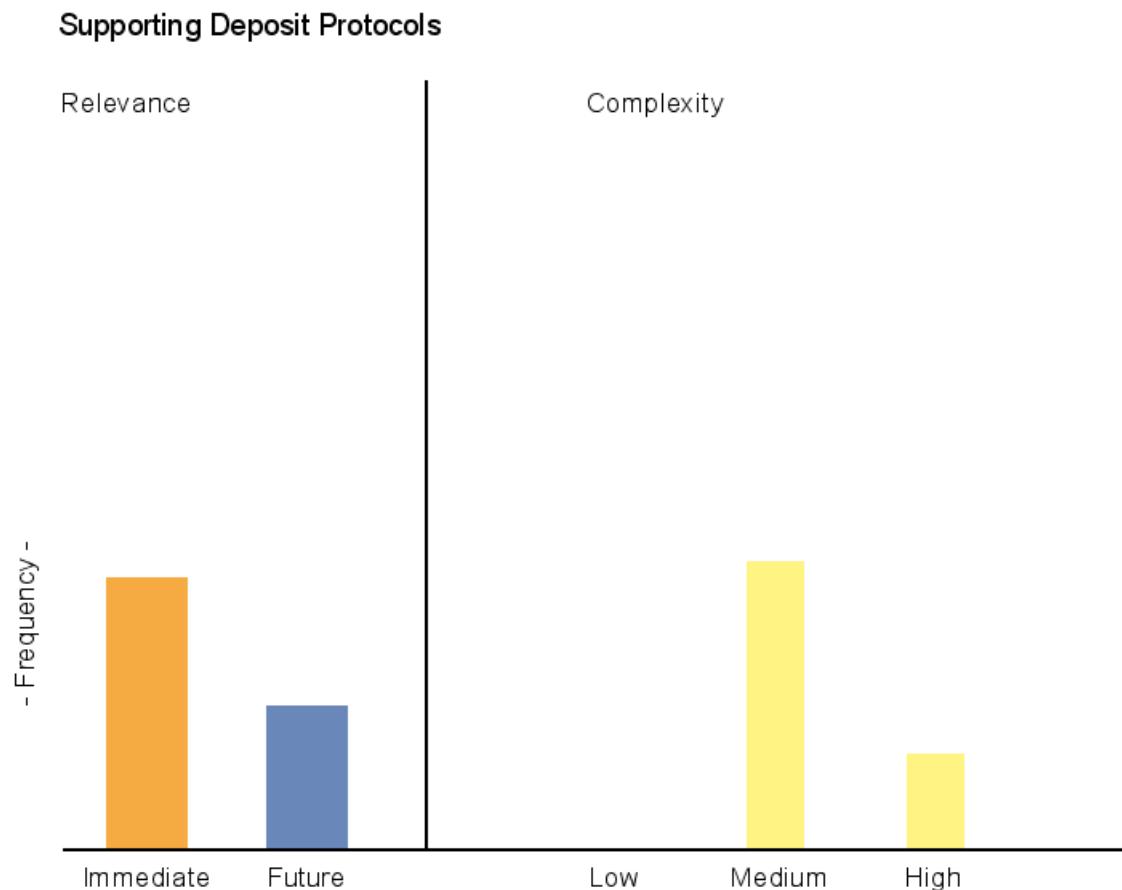
“Depends on dissemination of Guidelines and monitoring its adaptation and use”

“Especially relevant for large data aggregators like e.g. BASE, who are concurrently harvesting many repositories. An OAI element to convey this information in a yet unstructured way could be the description element of the Identify command of the OAI protocol.”

“You have to build them first”

### Issue: Supporting Deposit Protocols

As an example SWORD (Simple Web-service Offering Repository Deposit) is a standard that allows repositories to manage the deposit of content from external sources. Such protocols could be beneficial to handle the exchange of documents/objects in a bi-directional context.



### Comments

“Repositories as web services”

“Depends on systems. For AuthorClaim, not difficult.”

## Issues Scatter Diagrams

The scatter diagrams (one for the list of key aspects in total and one for each key aspect containing the corresponding issues) aim to make clear how relevant the issues are and how difficult the perceived implementation of the issues.

These diagrams aim to visualize the strategic aspects in the first instance: Immediate relevance ranks before future relevance and low complexity before high complexity. The two-dimensional position of the issue makes it more transparent which issue has which degree of strategic value. The top-left square contains the issues with immediate relevance and low complexity, thus should be the first choice for realization. Obviously on the other hand the topics in the bottom-right square contain the issues of high complexity with future relevance. This presentation should make the task of identifying the most efficient strategy for the different stakeholders.

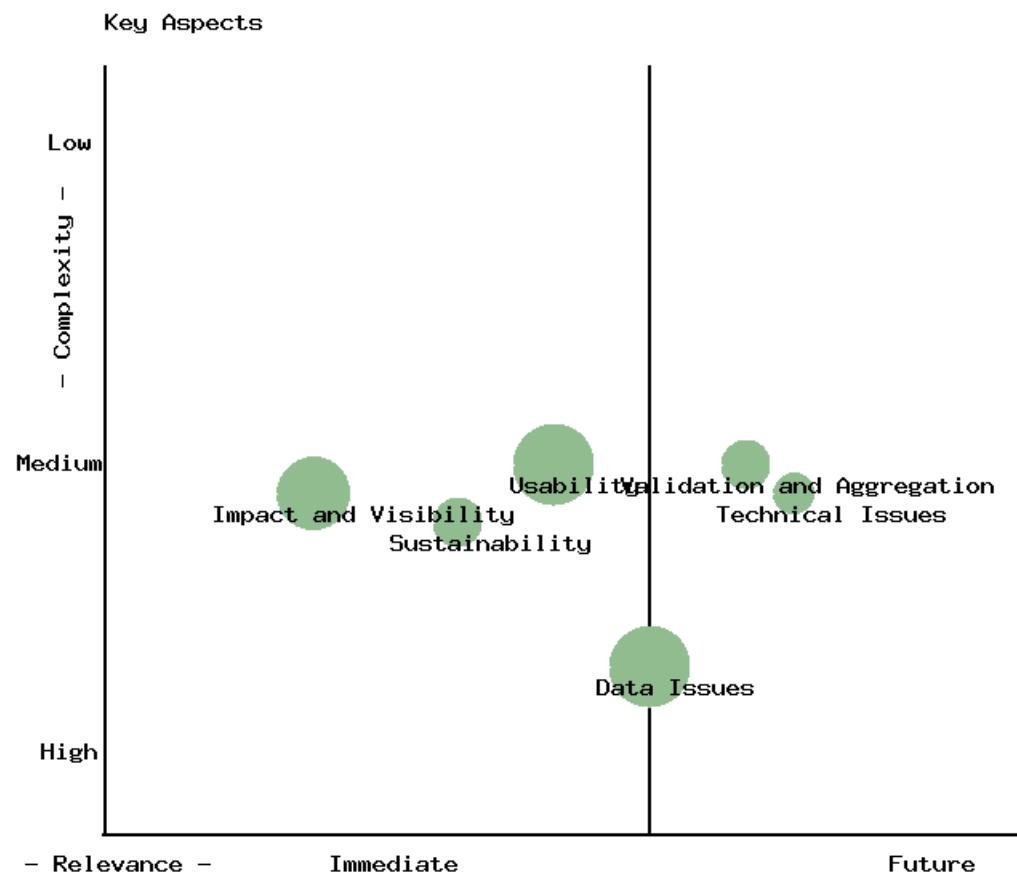
The diagrams are grouped by the key aspect topics with the axes relevance and complexity. The radius of the corresponding issue's circle is related to the degree of relevance. Accordingly focusing on the different key aspects brings up the following charts displaying their corresponding concrete issues in a scatter box with their according relevance, time frame of implementation and complexity grouped by their key aspect topics.

It could be an efficient strategy to look for the temporal relevance in the first instance and to take the complexity as second level into account. The scatter diagrams allow to evaluate the issues in groups based on this approach.

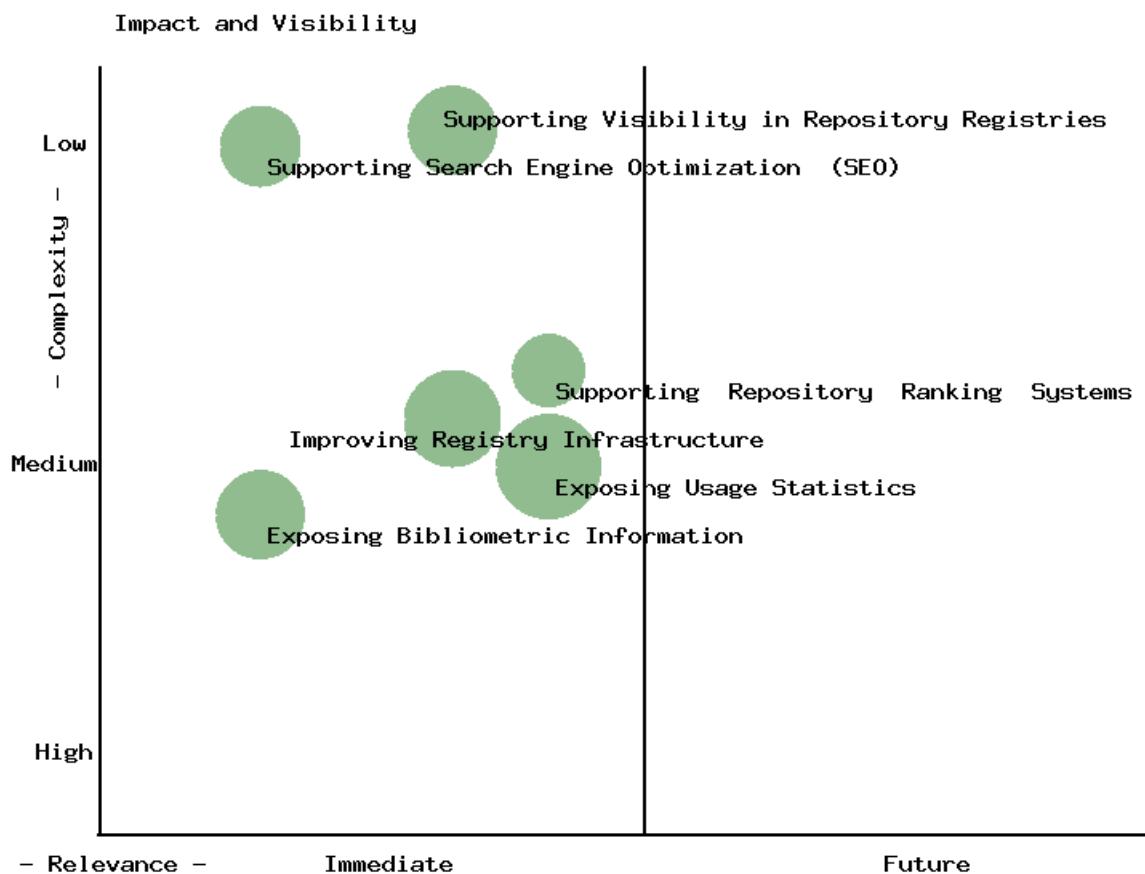
The first slide below merges the three dimensions listed above into one graph and shows the position of the key aspects (with their overall figures) among each other. Since these aspects include specific sub-issues of different type it delivers a more superficial overview of the relevance of the main topics.

## Overview Key Aspects

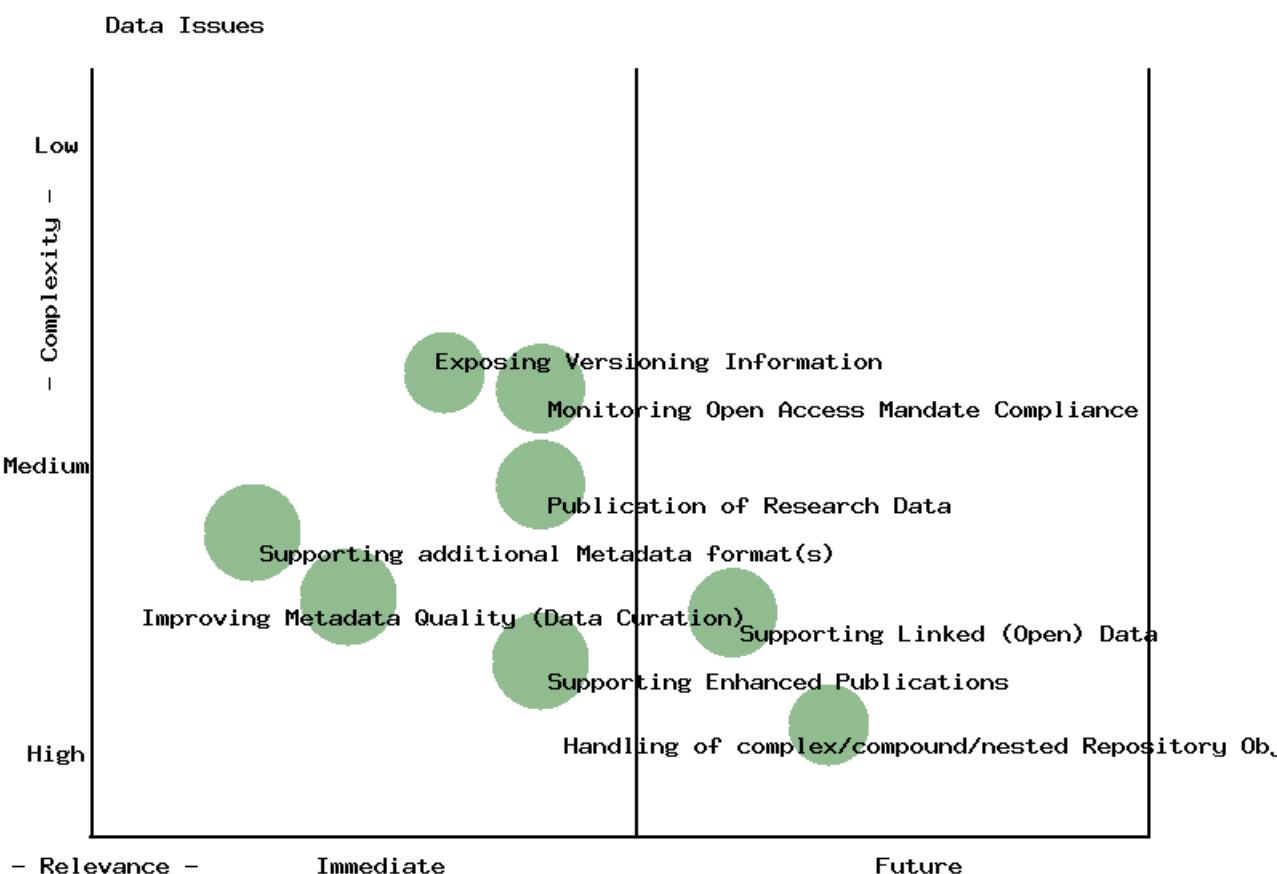
The overview diagram reflects the strategic estimation of the key aspect and not the average weight of the related specific issues.



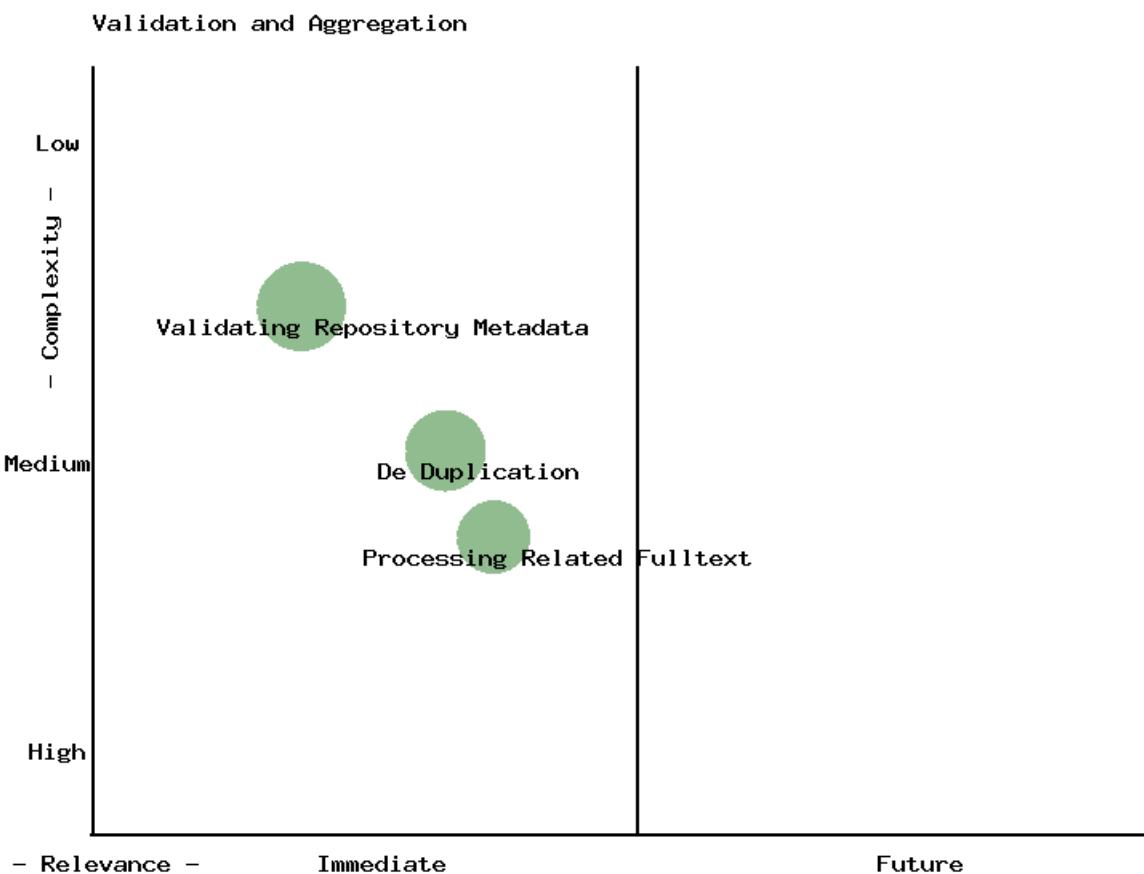
## Impact and Visibility



## Data Issues

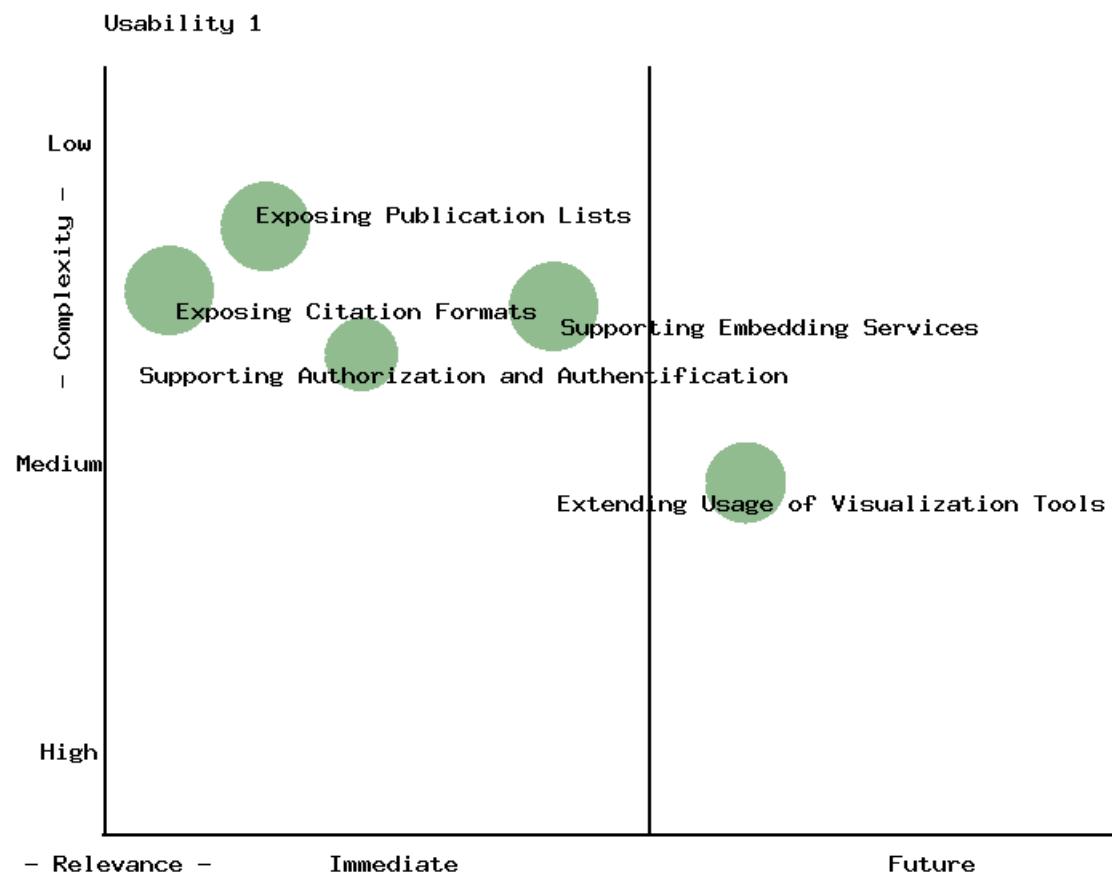


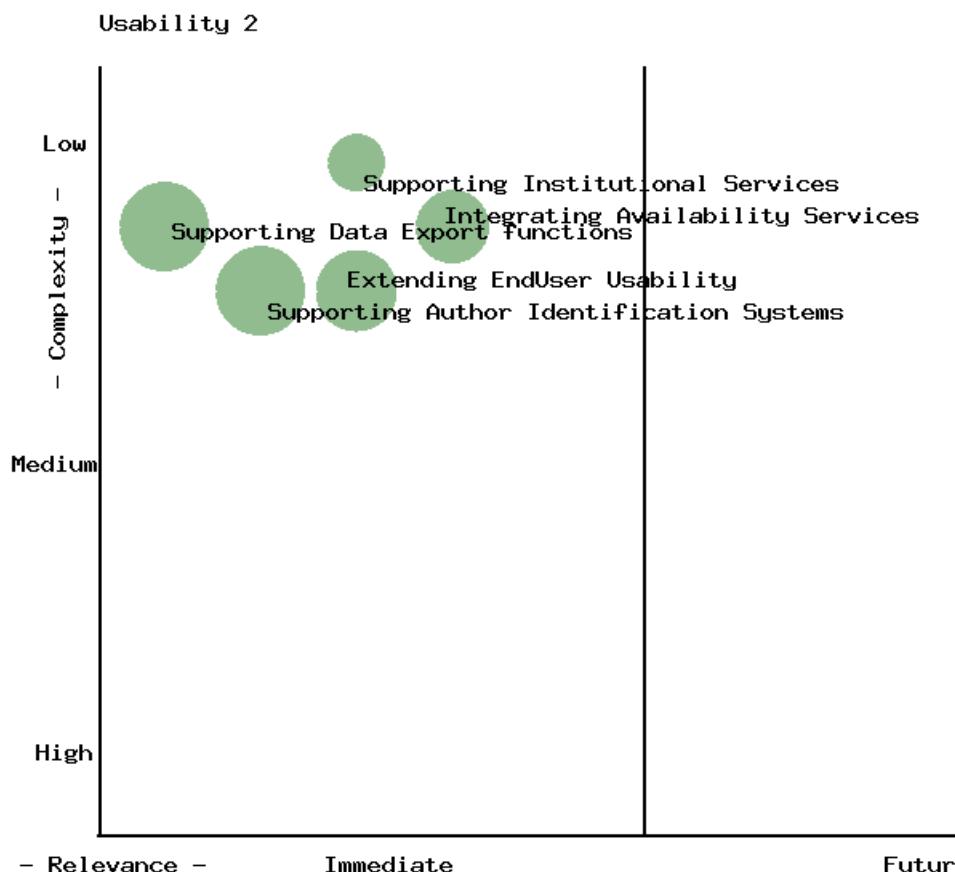
## Validation and Aggregation



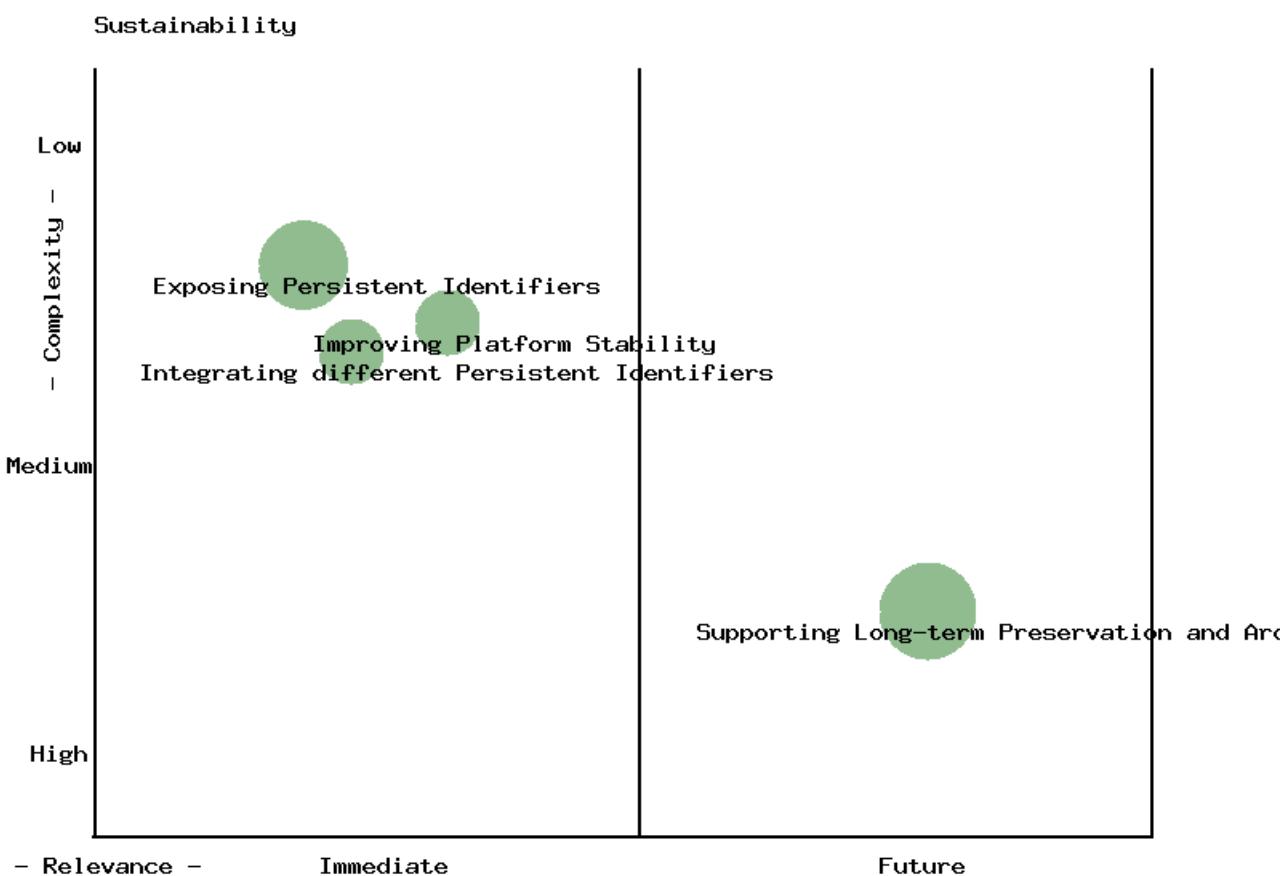
## Usability

The illustration of this aspect has been splitted into two slides since the single issues are at a very close range to each other.

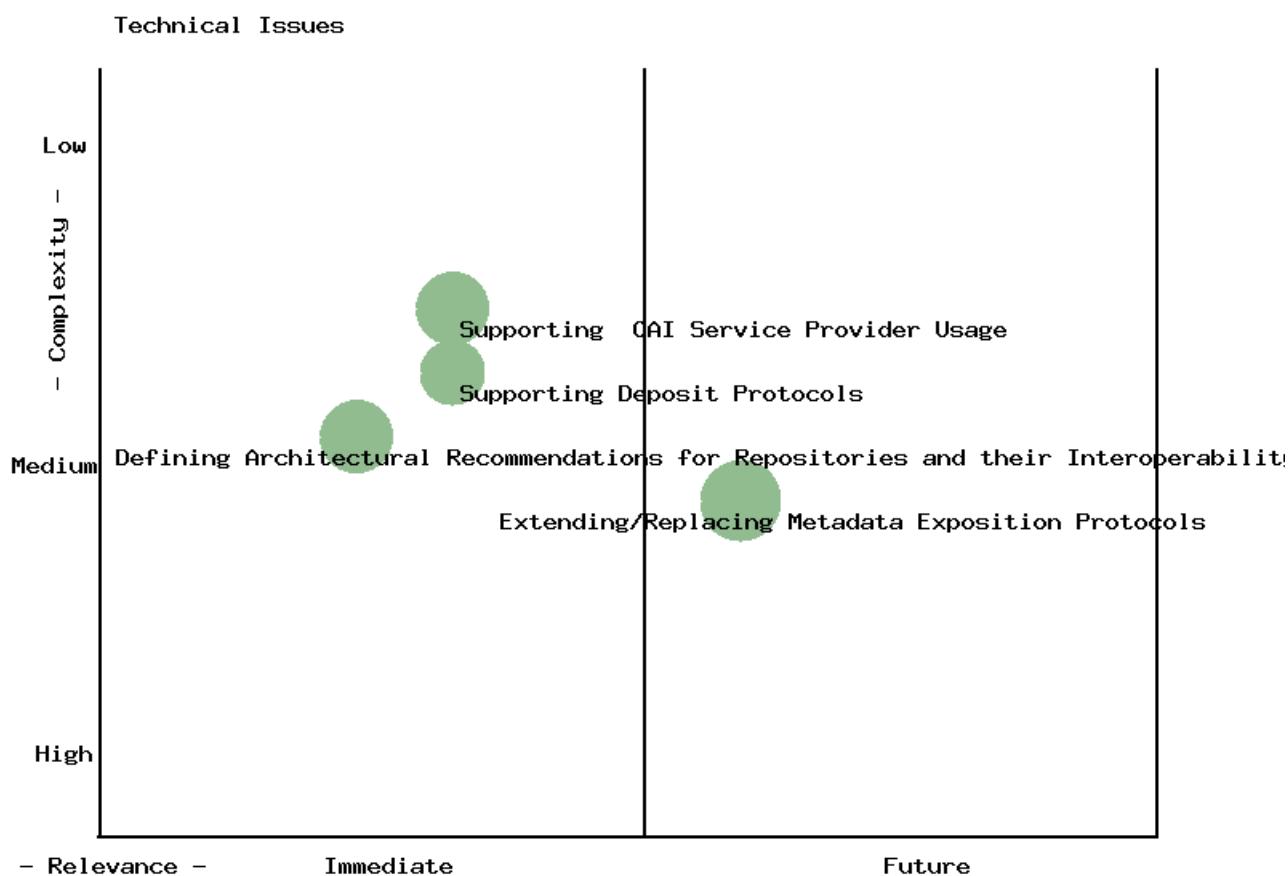




## Sustainability



## Technical Issues



## About COAR

COAR, the Confederation of Open Access Repositories, is a young association of repository initiatives launched in October 2009, uniting over 90 members and partners from 35 countries from throughout Europe, Latin America, Asia, North America and Australia. Its mission is to enhance greater visibility and application of research outputs through global networks of Open Access digital repositories. More information about COAR and its members is available on the COAR website [www.coar-repositories.org](http://www.coar-repositories.org).

## Contact Information

COAR Office

Katharina Müller, Head of Office

c/o Goettingen State and University Library

Platz der Gottinger Sieben 1

37073 Goettingen

Germany

Phone: +49 551 39-22215

Fax: +49 551 39-5222

[office@coar-repositories.org](mailto:office@coar-repositories.org)