Digital Libraries for the Preservation of Research Methods and Associated Artifacts

Raúl Palma
Poznan Supercomputing and Networking Center
Poznań, Poland
rpalma@man.poznan.pl

Piotr Hołubowicz
Poznan Supercomputing and Networking Center
Poznań, Poland
piotrhol@man.poznan.pl

Kevin Page
University of Oxford
Oxford, UK
kevin.page@oerc.ox.ac.uk

Oscar Corcho
Universidad Politécnica de Madrid
Madrid, Spain
ocorcho@fi.upm.es

Sara Pérez
Universidad Politécnica de Madrid
Madrid, Spain
zilphia@gmail.com

Cezary Mazurek
Poznan Supercomputing and Networking Center
Poznań, Poland
mazurek@man.poznan.pl

ABSTRACT

New digital artifacts are emerging in data-intensive science. For example, scientific workflows are executable descriptions of scientific procedures that define the sequence of computational steps in an automated data analysis, supporting reproducible research and the sharing and replication of best-practice and know-how through reuse. Workflows are specified at design time and interpreted through their execution in a variety of situations, environments, and domains. Hence it is essential to preserve both their static and dynamic aspects, along with the research context in which they are used. To achieve this, we propose the use of multidimensional digital objects (Research Objects) that aggregate the resources used and/or produced in scientific investigations, including workflow models, provenance of their executions, and links to the relevant associated resources, along with the provision of technological support for their preservation and efficient retrieval and reuse. In this direction, we specified a software architecture for the design and implementation of a Research Object preservation system, and realized this architecture with a set of services and clients, drawing together practices in digital libraries, preservation systems, workflow management, social networking and Semantic Web technologies. In this paper, we describe the backbone system of this realization, a digital library system built on top of dLibra.

1. INTRODUCTION

The preservation of digital information is one of the main topics addressed by archives and digital libraries. Digital preservation requires methods and technologies to ensure that digital information of continuing value remains accessible and usable [41]. This is particularly critical for the scientific community, where data is often time-dependent and cannot be recreated (satellite imagery, sensor data), or may be too costly or inconvenient to reproduce (e.g., MRI scans of a patient's organ).

Digital preservation is supported to some extent by digital library systems, which collect, manage and preserve digital content, with a measurable quality and according to codified policies [34]. Existing frameworks and approaches, such as OAIS [24] and Merritt [25], provide methodological grounding for describing, assessing and comparing the preservation capabilities of such systems.

Current systems are mainly focused on preserving content of a rather static nature, i.e., documents, images, datasets. However, research in data-intensive science, conducted in increasingly digital and online environments, has led to the emergence of other digital artifacts that have also a dynamic dimension (i.e. they are executable). One such artifact is the scientific workflow, an executable description of a scientific procedure, typically defining the sequence of computational steps in an automated data analysis. Scientific workflows are becoming widely used in many fields, as they are key in supporting reproducible science and the sharing and replication of best-practice and know-how through reuse [39]. Moreover, the research context of these scientific investigations is also dynamic in terms of their sharing, comments and lifecycle as they may evolve several times during their lifetime.

Therefore, the scientific community needs to curate and preserve not only data but also the associated processes that consume and generate that data, as well as essential metadata about the research context. The preservation of scientific workflows raises challenges [36] that deal with their dynamic aspects and their vulnerability to the volatility of the resources required for their execution, e.g., workflow decay. In addition to the workflow specification, preserving information about the data used and produced as a result of workflow execution, and the components that implement workflow steps, is essential to ensure its reproducibility. In order to deal with these challenges we require multidimensional, complex digital objects (called Research Objects [29]), which encapsulate essential information about experiments and investigations to facilitate the reusability, reproducibility and better understanding of scientific experiments. They comprise scientific assets, such as workflow models, provenance of their executions, datasets and other related resources, as well as their annotations. Moreover, we need appropriate technological support to allow scientific communities to preserve the static and dynamic aspects of their workflows, along with their enclosing research context, for subsequent sharing and reuse.

This paper presents the design and implementation of a specialized digital library system for the management, preservation, indexing and retrieval of semantic aggregations of workflows with related artifacts along with their enclosing research context. This system, built on top of dLibra digital library framework, realizes the backbone of a workflow-centric Research Object preservation infrastructure, supporting the evolution of these dynamic objects and providing specialized preservation features, such as monitoring of workflow decay. The design and implementation of the system considered the usage of open APIs for interoperability with other software components, and best practices in digital preservation according to standard preservation frameworks. The paper starts with a brief overview of related work (Section 2), followed by an introduction to the notion of Research Object (Section 3) and to the general software architecture of the preservation infrastructure (Section 4). Next, we present the implementation (Section 5) of the digital library system, including an overview of the APIs it implements and current client applications, including the RO portal, myExperiment portal and a command line tool. Then, we discuss the system support for managing and preserving the dynamic aspects of research objects (Section 6). Finally, we report an analysis of our implementation against standard preservation frameworks (Section 7) and we conclude in Section 8.

2. RELATED WORK

Relevant digital library (DL) systems include Greenstone [8], Digital Library eXtension Service [22], gCube DL management system [7], JeromeDL [9], dLibra [3], etc. These systems are mostly used for the publication of documents and other forms of non-executable digital content. Similarly, well-known repositories like Fedora [6], DSPACE [4], or EPrints [5], are used in scholarly environments to create institutional repositories to manage documents and collections. Some other systems, like LOCKSS [10], are decentralized and rely on lots of copies to keep stuff safe, acknowledge that even scholarly content may change and multiple versions may be produced, and focus on preserving and providing access to the original content without managing their evolution. Our system extends dLibra with the notion of digital objects linked with semantic relations, addressing their evolution, and with specialized preservation features, such as monitoring of workflow decay.

Social scientific networking sites (aka e-laboratories) bring together researchers and resources in a Web 2.0 context. Examples are myExperiment [37], SysMO-DB [21] and MethodBox [11]. These sites focus on the storage, sharing and community driven collaboration around scientific methods and processes. Similarly, the Open Science Framework [14] provides an environment for documenting, archiving and sharing scientific projects comprising different research materials - tools, scripts, methods, measures and data. Some of these infrastructures also provide additional features, such as means for organizing and commenting the resources and even some basic versioning support. However, in general they do not provide means for describing the complex relationships between these related resources, or their dynamic aspects (e.g., runs, execution logs, provenance traces, etc.), including the evolution of the scientific investigations they represent. Moreover, they don't address challenges associated with the preservation of dynamic objects, such as error checking, disaster recovery, migration, monitoring and notification of decay or other issues compromising the reproducibility of experiments. Our system can be integrated with such social sites to complement each other.

3. OVERVIEW: WORKFLOW-CENTRIC RESEARCH OBJECTS

Our workflow-centric Research Object model, introduced in [31], provides the means for capturing and describing aggregations of scientific assets and their annotations, facilitating the reusability, reproducibility and better understanding of the scientific investigations they represent. These objects comprise elements describing the way research findings are produced since the formulation of a hypothesis, the design and execution of the experiments, the analysis of results and the conclusion makings, including data used and produced, methods employed (encoded for instance by one or several scientific workflows), provenance and setting information, people involved and annotations about these resources.

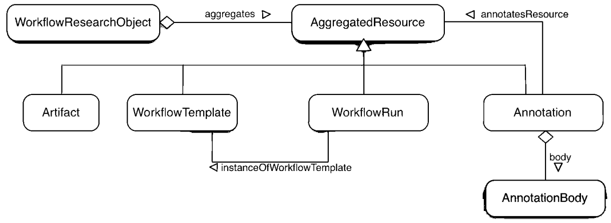


Figure 1: Workflow-centric research object as an aggregation of resources

The model consists of a suite of ontologies: the core ro ontology, which provides basic structure for the description of aggregated resources and annotations on those resources; wfdesc, for the description of workflows that provides an abstraction layer over different workflow systems; wfprov, for describing provenance of workflow results and executions; and roevo, for describing the evolution of research objects, including the different stages during their lifecycle, the corresponding versions of these objects and their aggregated resources along with their associated changes. These ontologies were built upon existing vocabularies as much as possible, including OAI ORE (Object Exchange and Reuse) [13] for specifying aggregation of resources, the Annotation Ontology [35] to support the annotation of research objects, their constituent resources, as well as their relationship, and the PROV Ontology to represent provenance information. wfprov and roevo, for instance, build upon PROV Ontology [15] and extend it with terms that capture complementary provenance information about research objects and their aggregated resources, i.e., provenance of workflow results and provenance about their evolution and versioning.

Figure 1 illustrates a coarse-grained view of a workflow-centric research object, which aggregates a number of resources: a workflow template, which defines the workflow; workflow runs obtained by enacting the workflow template; other artifacts, e.g., a paper that describes the research, datasets used in the experiments, etc.; and annotations describing the aforementioned elements and their relationships. A complete description of the Research Object model, including examples and links to their OWL versions can be found in its website [16], and more technical details at [32].

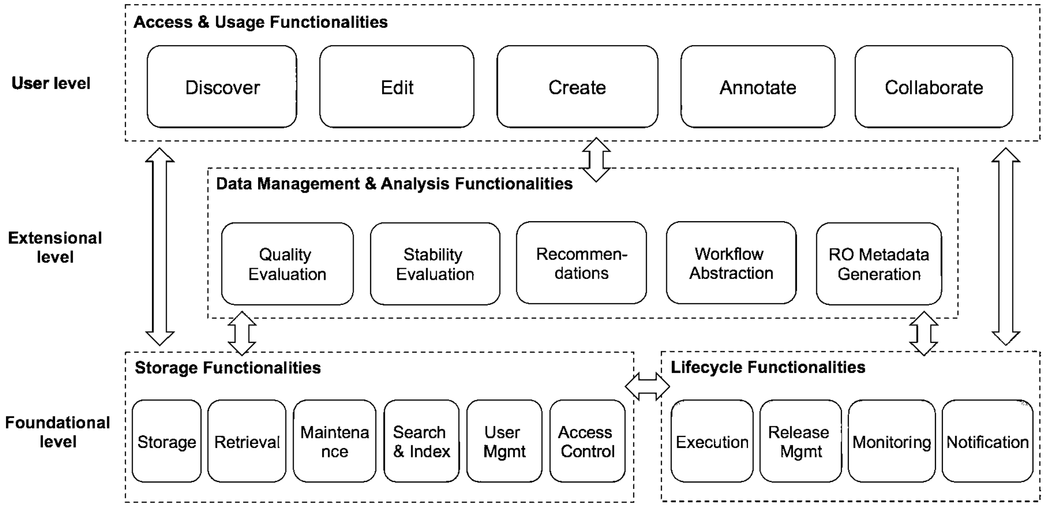


Figure 2: High-level architecture overview

4. SOFTWARE ARCHITECTURE FOR RESEARCH OBJECT PRESERVATION

The architecture, introduced in [43] and described in detail in [38], specifies functionalities required by a workflow-centric Research Object preservation infrastructure, along with a set of functional entities for their grouping and categorization, enclosed in a three layer framework (foundational, extensional, clients) - see Figure 2. The architecture is compliant with preservation standards (see Section 6) and best practices from Web and Linked Data approaches. We have employed Linked Data [33] and REST [44] as our main architectural principles. We argue that these approaches, both centered on the notion of resources, are complementary: both are based on the use of URIs, Web-style linking, and HTTP as the transfer mechanism. Linked Data focuses on the linking of common resources in an RDF representation, while REST uses the navigation between resources to progress application state (such that client-server interaction is stateless). In combination this leads to a design method driven by the identification of resources and their representations [42]. Moreover, in identifying the functional entities, we aim to aid designers of future systems by providing a more precise set of terms and concepts for use as reference model, and to ease alignment of our architecture with standards like the OAIS reference model.

The primary focus of the architecture is interoperability and compatibility between different software components. Thus, we provide simple, lightweight, and adaptable APIs to the identified functionalities.Examples of these APIs include the RO API, the RO Evolution API, Notification API, etc.[1] These APIs interoperate through the Research Object and related models, the data structures that encode the concepts and relationships of information.

The foundational layer of the architecture includes services from two functional entities: Storage and Lifecycle, which together address static and dynamic aspects of Research Objects (ROs):

Storage entity prescribes the underlying services for the storage, maintenance, search and retrieval of RO resources, ensuring appropriate levels of protection. They provide capacities to the preservation infrastructure, concerned mainly with the static aspects of preservation, e.g., adding resources to permanent storage, managing internal data models, maintaining the RO resources associated semantic metadata, indexing, searching and protection of resources, basic versioning mechanisms, etc.

Lifecycle entity addresses the dynamic nature of RO resources. It prescribes services for the execution of RO resources, such as workflows, the management of RO evolution, including versioning and curation activities, and the provision of monitoring and notification mechanisms to ensure the correct preservation of ROs.

On top of this layer, the architecture prescribes an (i) extensional layer of *Data Management & Analysis* services that generate, maintain and provide access to added-value data derived from or related to RO resources (e.g., quality-related information, recommendations and metadata extracted); (ii) an *Access & Usage* layer that enable users to interact with research objects.

5. RO DIGITAL LIBRARY

The foundational service to preserve workflow-centric research objects is the Research Object Digital Library (RODL), which realizes the Storage and Lifecycle functionalities described in Section 4. It collects, manages and preserves aggregations of workflows and related objects and annotations, packed into research objects.

5.1 The interfaces

The main interface of RODL is a set of REST APIs, being the two primary ones the RO API [17] and the RO Evolution API [18].

The RO API, also called the RO Storage and Retrieval API, defines the formats and links used to create and maintain research objects in the digital library. It is aligned with the RO model, hence recognizing concepts such as aggregations, annotations and folders. The RO model ontology [32] is used to specify relations between different resources. The RODL supports content negotiation for metadata, including formats like RDF/XML, Turtle and TriG.

The RO Evolution API defines the formats and links used to change the lifecycle stage of a research object, most importantly to create an immutable snapshot or archive from a mutable live research object, as well as to retrieve the evolution provenance of a research object. The API follows the RO evolution model [40].

Additionally, RODL provides a SPARQL endpoint, a Notification API [12], a Solr REST API, an OpenSearch API, and a custom User Management API [23].

5.2 The implementation

[1] Detailed information of APIs is available at `http://www.wf4ever-project.org/wiki/display/docs/Wf4Ever+service+APIs`

One of the main design challenges related to the implementation of RODL was the need to support live, dynamically changing research objects as well as immutable snapshots that are intended for longterm preservation. The RODL has a modular structure that comprises access components, longterm components and the controller that manages the flow of data (see figure 3). For immutable research objects, they are stored in the longterm preservation repository once created. The live ROs, on the other hand, are pushed asynchronously after every change or periodically.

The access components are the storage backend - dLibra [3] - and the semantic metadata triplestore. dLibra provides file storage and retrieval functionalities, including file versioning and consistency checking. It has a built-in text search engine and it manages users and controls their access rights. It allows organizing stored objects into hierarchical structures and associating metadata at the level of object aggregations. It is also possible to use a built-in module for storing research objects directly in the filesystem.

The semantic metadata are additionally parsed and stored in the triplestore backed by Jena TDB [27]. Jena TDB is an actively developed RDF store implementation, which provides good support for transactions, querying, cacheing and using named graphs. The use of a triplestore helps in RODL internal data processing and offers a standard query mechanism for RODL clients. It also provides a flexible mechanism for storing metadata about any component of a research object that is identiable via a URI, which apart from workflows and other resources, may include parts of workflows or external resources (e.g. web services, data sources).

The longterm preservation component is built on dArceo [2]. dArceo stores objects and monitors their quality, alerting administrators if necessary. Standard monitoring activities include file format decay alerts and fixity checking but can be enhanced using a plugin mechanism. In case of RODL, dArceo monitors the quality of research objects by calling the Checklist Evaluation and Stability Services [1, 20]. If a change in quality is detected, notifications are generated as Atom feeds in compliance with the Notification API mentioned above. This helps detect and prevent workflow decay which occurs when an external resource or service used by the workflow becomes unavailable or is otherwise behaving differently.

dArceo allows defining migration plans to perform a batch update of resources from one format to another, if necessary. In case of workflows, this may be applied for instance when a flat Taverna t2flow format should be converted to a complex scufl2 format (which, notabene, uses the RO model similarly to research objects). Other case could be a batch update of workflows that depend on a malfunctioning external resource. Objects in dArceo can be stored on a range of backends, including specialized preservation repositories such as the Platon service [28], storing data in geographically distributed copies and guaranteeing consistency.

An RODL running instance, with more than 1300 ROs, is available at `http://sandbox.wf4ever-project.org/rodl/`.

5.3 RODL clients

The reference client of RODL is **the RO Portal**, developed alongside RODL to expose all available functionalities. It is a web application running at `http://sandbox.wf4ever-project.org/portal`. Its main features are research object exploration and visualization; it also allows to create user accounts in RODL and generate access tokens for other clients. The RO Portal uses all APIs of RODL. Figure 4 shows the main view in RO portal of an exemplary research object for sharing an experiment protocol to enhance its reproducibility[2]. The development version of **myExperiment** [37] (`http://alpha.myExperiment.org`) uses RODL as a backend for storing packs. It uses the RO API. Finally, the **RO Manager** [19] is a command line tool that is primarily used to manage a research object stored on a local disk. It allows to push a research object

[2] A more detailed description of this exemplary research object is at `http://www.researchobject.org/examples/58-2/`

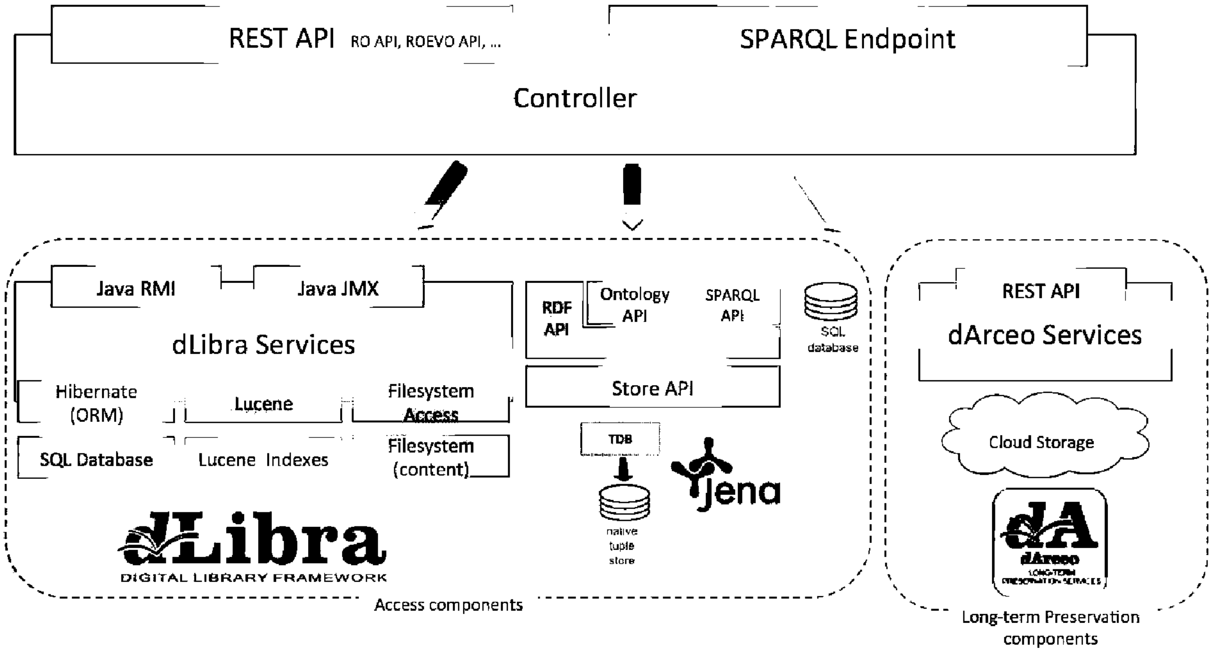


Figure 3: Research Object Digital Library internal component diagram

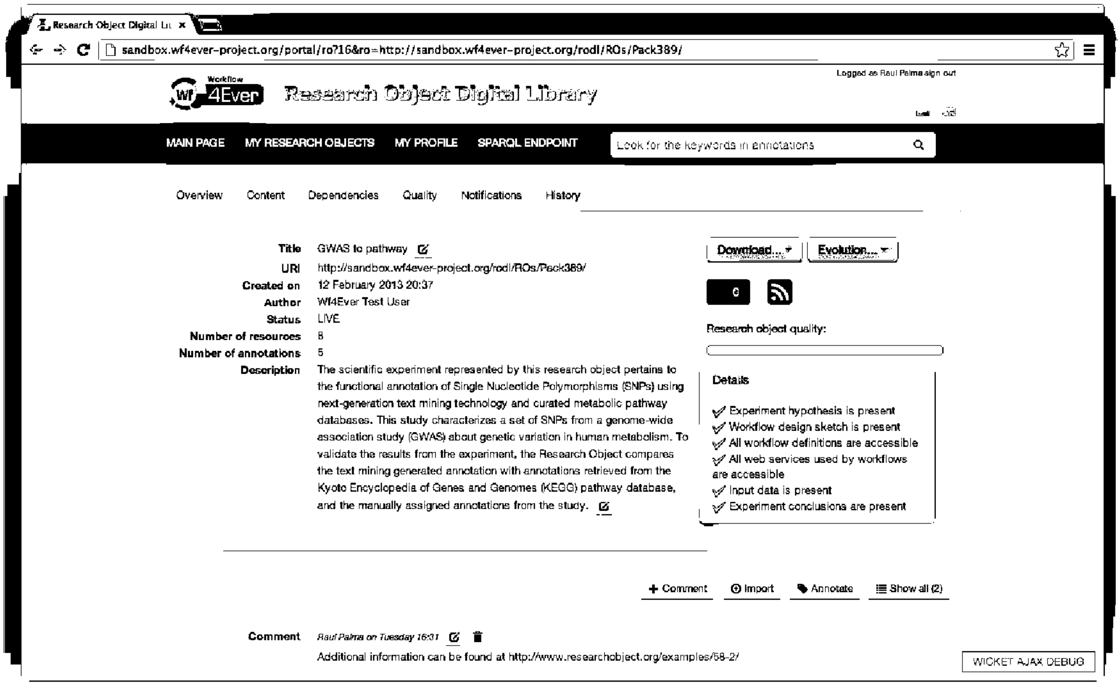
to RODL via the RO API, as well as convert it into a snapshot in RODL.

6. MANAGING RESEARCH OBJECT EVOLUTION

As aforementioned, one of the key features of RODL is its support for managing and preserving the dynamic aspects of research objects. This is achieved by providing the means for capturing the lifecycle of a research object, that is, the states that it transition since its conception until its archival, including the provenance information associated with the research object and the resources it aggregates, particularly workflows and other executable processes. For instance a workflow may have several runs associated, each using and producing different artifacts. The provenance of the results of a workflow run may include the description of execution steps, responsible actors, along with the input and outputs artifacts of the workflow and its subprocesses. Similarly, provenance information about the research object evolution is related to their transformation into a new state, capturing the versioning information. This information includes when the transformation was done, from which research object this new state was derived, as well as how, when and by whom the research object was modified keeping track of the contributions and attributions in the research object and its aggregated resources (e.g., the source or entity contributing to the artifact). Additional provenance information may include justifications, assumptions or other user provided annotations that facilitate the understanding of the scientific investigation.

We illustrate RODL support for the evolution of research objects with an exemplary lifecycle. A research object normally starts its life as an empty Live research object, with a first design of the experiments to be performed. It is then incrementally filled by aggregating datasets, documents, workflows encoding scientific methods, and other related resources. These resources may be generated from scratch or by reusing and repurposing existing resources, keeping the record of contributions and attributions. This information may be generated automatically by RODL (e.g., the original source), or provided by the user (e.g., derived from). While working with the Live research object, its resources may be changed at any point in time, new resources may be added, other may be removed, and they may be annotated. Moreover, executable resources such as workflows may be run several times using different inputs and producing different outputs, generating provenance information of the results that may be recorded in RODL.

At several points in time this Live research object gets copied and kept into a research object Snapshot, which aims to reflect the status of the research object at a given point in time. Such a Snapshot may be useful to release the current version of the research outcome of an experiment, submit it to be peer reviewed or to be published (with the appropriate access control mechanisms), share it with supervisors or collaborators, or for acknowledgement and citation purposes. For each Snapshot, RODL automatically captures the versioning information (e.g., previous version), when this Snapshot was created, from which Live object is derived, by whom and how it was changed from the previous version (e.g., the set of changes). Changes and versioning information may be captured at the level of the research object and its aggregated resources. Snapshots have their own identifiers, and may be preserved in RODL, since it may be useful to be able to track the evolution of the research object over time, so as to allow, for example, retrieval of a previous state of the research object, reporting to funding agencies the evolution of the research conducted, etc. Additionally, Snapshots may be monitored by RODL to record quality-related information of the research object at certain point in time, providing


Research Object Digital Li
sandbox.wf4ever-project.org/portal/ro?16&ro=http://sandbox.wf4ever-project.org/rodl/ROs/Pack389/
Logged as Raul Palma sign out
Workflow 4Ever
Research Object Digital Library
MAIN PAGE
MY RESEARCH OBJECTS
MY PROFILE
SPARQL ENDPOINT
Look for the keywords in annotations
Overview
Content
Dependencies
Quality
Notifications
History
Download...
Evolution...
0
+ Comment
Import
Annotate
Show all (2)
WICKET AJAX DEBUG

Title GWAS to pathway

URI http://sandbox.wf4ever-project.org/rodl/ROs/Pack389/

Created on 12 February 2013 20:37

Author Wf4Ever Test User

Status LIVE

Number of resources 8

Number of annotations 5

Description The scientific experiment represented by this research object pertains to the functional annotation of Single Nucleotide Polymorphisms (SNPs) using next-generation text mining technology and curated metabolic pathway databases. This study characterizes a set of SNPs from a genome-wide association study (GWAS) about genetic variation in human metabolism. To validate the results from the experiment, the Research Object compares the text mining generated annotation with annotations retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database, and the manually assigned annotations from the study.

Research object quality:

Details

- ✓ Experiment hypothesis is present
- ✓ Workflow design sketch is present
- ✓ All workflow definitions are accessible
- ✓ All web services used by workflows are accessible
- ✓ Input data is present
- ✓ Experiment conclusions are present

Comment *Raul Palma on Tuesday 16:31*

Additional information can be found at http://www.researchobject.org/examples/58-2/

Figure 4: The Research Object Portal

useful information of how its quality has changed through its life.

At some point in time, the research object may get published and preserved in RODL, in an Archived research object. It has a permanent identifier and may be created by copying completely the Live research object, or it may be the result of some filtering or curation process where only some parts of the information available in the aggregation are actually published for others to reuse. Later on, an Archived research object can be used as a starting point for a new research work, e.g., by repurpose it or its parts, in which case it is used as the base to create a new Live research object.

7. COMPLIANCE WITH PRESERVATION STANDARDS

The design and implementation of our preservation infrastructure was also driven by a twofold approach for testing its compliance against preservation standards: a) mapping our concepts and functionalities to the OAIS standard, analyzing the information model and functional entities roles, services and functions defined by OAIS and checking whether they are present or can be mapped in the architecture; similar to the approach taken by the National Archives (TNA) and the UK Data Archive (UKDA) [30]; b) the Merritt micro-services curation approach developed by the UC3 (University of California Curation Center) [25][26]. UC3 microservices have been aligned to the "The DCC curation lifecycle model". They can be grouped into four categories that provide incrementally increasing levels of preservation assurance and curation value. This approach has been used to analyze our preservation services.

7.1 Alignment to the OAIS reference model

The OAIS (Open Archival Information System) reference model, ISO 14721-2012 standard, defines the basic functional components of a system dedicated to the long-term preservation of digital information, identifies and describes the entities which constitute the OAIS environment where the archive operates, details the key internal and external system interfaces which allow the interaction with these entities, and characterizes the information objects managed by the system. The reference model also enumerates a set of minimum requirements an archival system is expected to meet.

Following an approach similar to [30], first we mapped our objects to the concepts of the OAIS information model, and then analyzed how each OAIS function is covered by our infrastructure. Regarding the OAIS information model, the information objects or information packages managed by an archival system comprise content information (the target information to be preserved) and preservation description information (metadata necessary to identify and understand the environment in which the content was created). Information packages are categorized into Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP). SIP is the package sent to an OAIS by a Producer, which is transformed into one or more AIP for preservation. When requested by a Consumer, OAIS provides all or part of the AIP in the form of a DIP.

The concept of Information Package in OAIS can be roughly mapped to a Research Object, as depicted in figure 5. Research Objects, similar to Information Package, are conceptual containers (aggregations) of Resources (Content Information) and annotations about them (metadata), including provenance, context and reference information. The package submitted to the RO infrastructure (SIP) can be basically individual resources (e.g., workflows, datasets) or simple aggregation of resources. Within the RO infrastructure, these resources are transformed into Research Objects, generating metadata during the transformation regarding the provenance, context, semantic relations, etc. Similarly, one workflow submitted to the infrastructure may generate a Research Object, which will include the original workflow, plus a workflow bundle (a new workflow format), metadata extracted from the workflow regarding its structure (e.g., inputs, outputs), and other metadata generated. When they are considered to be immutable, ROs are sent to the preservation infrastructure (AIPs). Furthermore, in response to a request, the infrastructure can return (DIP) the whole Research Object, in different representations (e.g. as a ZIP archive, HTML page, metadata), or it can provide individual resources within the Research Object (e.g. workflows, datasets).

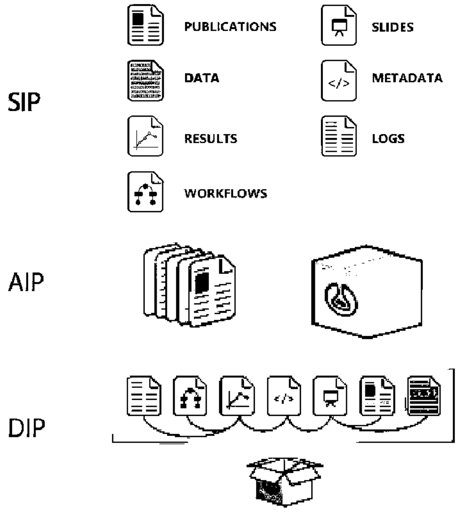


Figure 5: RO components and OAIS Information Packages

Once information packages were clarified, we proceeded with the OAIS functions analysis. The OAIS functional model comprises six functional entities (ingest, archival storage, data management, administration, preservation planning and access) that fulfill the OAIS role of preserving information and making it available to a designated community. The results [38] show a good coverage of these functions by RODL, in particular regarding the technical aspects in Ingest, Archival Storage, Data Management and Access.

7.2 Alignment to Merritt microservices

Digital preservation is implemented through complex procedures. Breaking these life-cycle processes down into smaller manageable tasks is one of the rationales for providing services that address a clear issue and solve it in an efficient way. An approach that emerged a few years ago and represents a step away from integrated digital archive systems is the one of microservices. These allow to flexibly combining specialized solutions for preservation depending on the requirements of the institution [45].

The University of California Curation Center[3] has defined a micro-services approach to curation that could serve as a basis for our purposes. These microservices work as a general list of requirements for preserving digital objects, such as Research Objects. By identifying how each of them was implemented we could check to what extent Wf4Ever met the general preservation requirements.

These services can be seen as a progression depending on the level of preservation and curation value they offer. The two first groups or levels, Protection (providing security through redundancy) and Interpretation (maintaining meaning through descrip-

[3] UC3 - `http://www.cdlib.org/services/uc3/`

tive context) are considered *preservation* functionalities as they ensure the integrity of the object, the content state and context. The two last groups, Application (facilitating utility) and Interoperation (adding value) are considered *curation* services, as they focus on the content of the digital object and their goal is to maintain the value and usefulness / service of their content.

A complete description of the analysis against the Merritt catalogue of microservices can be found at [38], depicted in tables 1 and 2.

Table 1: Protection services

UC3 microservice	*Wf4Ever implementation*
Identity service	URIs are used for identifying all entities.
Storage service	RODL is used for storing Research Objects. External resources aggregated by the Research Object are only referenced.
Fixity service	RODL uses checksums at file level.
Replication service	RODL provides replication services. Contents are replicated in dArceo, which are stored on a secure storage platform offering geographical replication.

Table 2: Interpretation services

UC3 microservice	*Wf4Ever implementation*
Inventory service	RO model permits annotations, which are used for arbitrary and user-generated metadata. RODL generates annotations for the evolution information and lets users submit their own annotations. Additionally, RODL generates basic descriptive metadata, e.g., creator, creation date, etc.
Characterization service	RODL stores and publishes metadata as RO annotations, based on the RO model.

8. CONCLUSIONS

We have designed and implemented a digital library system for the management, preservation, indexing and retrieval of research methods and related artifacts, addressing the requirements of "executable" content and respecting standards and practice in the digital library community. In order to do so, we introduced the notion of complex and multidimensional digital objects, called research objects, that encapsulate these artifacts along with the context information of the scientific investigation they represent. In particular, research objects comprise essential information relating to experiments and investigations that facilitate the reusability, reproducibility and better understanding of scientific experiments. The Research Object model has been implemented as a suite of lightweight ontologies, building upon existing work from related communities.

The digital library system realizes the backbone services of a software architecture for the preservation of workflow-centric research objects. The architecture prescribes a set of REST APIs to the identified functionalities, which are built around the research object model and its extensions.

The implementation of the digital library system extends dLibra digital library framework with the notion of research objects, associated semantic relations, addressing the evolution of these dynamic objects, and providing specialized preservation features, such as monitoring of workflow decay and other issues that may affect the quality of the research object. In doing so, we incorporated a triplestore backed by Jena TDB and a long-term preservation component built on dArceo system. The main interface to the digital library system is the set of REST APIs it implements, including the core APIs prescribed by the architecture, although it also provides a SPARQL endpoint. Users interact with the system via client applications. Currently, there are three different client applications, which support some or all of the APIs implemented by the digital library system. The reference client, RO portal, developed alongside the digital library system exposes all available functionalities. The development of myExperiment portal can use the system for storing packs, and RO manager command line tool can push local research objects into the system or pull them locally. An instance of the research object digital library, running in the sandbox of Wf4Ever project, contains more than one thousands research objects, many of them created by migrating and transforming workflows in myExperiment portal.

We have also described the approach followed during the design and implementation of our infrastructure for testing its conformance against preservation standards, making sure we are respecting standards and following best practices in digital preservation. In particular, we have (i) aligned our objects to the concepts of the OAIS information model, and analyzed how each OAIS function is covered by our infrastructure; (ii) analyzed preservation services of our realization against the Merritt catalogue of microservices.

9. ACKNOWLEDGEMENTS

This research is supported by the Wf4Ever EU project (http://www.wf4ever-project.org, FP7-270129). We acknowledge Gema Bueno for her early contributions on the compliance with preservation standards, and Graham Klyne and Stian Soiland-Reyes for their contributions to APIs.

10. REFERENCES

[1] Checklist Evaluation API. `http://wf4ever-project.org/wiki/display/docs/RO+checklist+evaluation+API`. [Online; accessed 20-May-2013].

[2] dArceo. `http://dlab.psnc.pl/darceo/`. [Online; accessed 20-May-2013].

[3] dLibra. `http://dlab.psnc.pl/dlibra/`. [Online; accessed 20-May-2013].

[4] DSPACE. `http://www.dspace.org/`. [Online; accessed 20-May-2013].

[5] EPrints. `http://www.eprints.org`. [Online; accessed 20-May-2013].

[6] Fedora Commons. `http://www.fedora-commons.org`. [Online; accessed 20-May-2013].

[7] gCube. `http://www.gcube-system.org/`. [Online; accessed 20-May-2013].

[8] Greenstone. `http://www.greenstone.org`. [Online; accessed 20-May-2013].

[9] JeromeDL. `http://www.jeromedl.org/`. [Online; accessed 20-May-2013].

[10] LOCKSS. `http://www.lockss.org/`. [Online; accessed 20-May-2013].

[11] MethodBox. `https://www.methodbox.org`. [Online; accessed 20-May-2013].

[12] Notification API. `http://wf4ever-project.org/wiki/display/docs/Notification+API`. [Online; accessed 20-May-2013].

[13] Open Archives Initiative Object Reuse and Exchange. `http://www.openarchives.org/ore/1.0/`. [Online; accessed 20-May-2013].

[14] Open Science Framework. `http://www.openscienceframework.org`.

[Online; accessed 20-May-2013].

[15] PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/. [Online; accessed 20-May-2013].

[16] Research Object model website. http://www.researchobject.org/. [Online; accessed 20-May-2013].

[17] RO API. http://wf4ever-project.org/wiki/display/docs/RO+API+6. [Online; accessed 20-May-2013].

[18] RO Evolution API. http://wf4ever-project.org/wiki/display/docs/RO+evolution+API. [Online; accessed 20-May-2013].

[19] RO Manager. https://github.com/wf4ever/ro-manager/. [Online; accessed 20-May-2013].

[20] Stability Evaluation API. http://wf4ever-project.org/wiki/display/docs/Stability+Evaluation+API. [Online; accessed 20-May-2013].

[21] SysMO-DB. http://www.sysmo-db.org. [Online; accessed 20-May-2013].

[22] The University of Michigan Digital Library eXtension Service. http://www.dlxs.org/. [Online; accessed 20-May-2013].

[23] User Management API. http://wf4ever-project.org/wiki/display/docs/User+Management+2. [Online; accessed 20-May-2013].

[24] Reference model for an open archival information system (OAIS) : Recommendation for space data system standards : CCSDS 650.0-B-1. Technical report, Jan. 2002.

[25] Merritt: An Emergent Approach to Digital Curation Infrastructure. Technical report, 2010.

[26] UC3 Curation Foundations. https://confluence.ucop.edu/download/attachments/13860983/UC3-Foundations-latest.pdf, 2010. [Online; accessed 20-May-2013].

[27] Apache Jena. http://jena.apache.org/, 2013. [Online; accessed 20-May-2013].

[28] PLATON - Science Services Platform. Technical report, PIONIER Consortium, 2013. [Online; accessed 20-May-2013].

[29] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, and E. Al. Why linked data is not enough for scientists. *Future Generation Computer Systems (FGCS)*, 2011.

[30] H. Beedham, J. Missen, M. Palmer, and R. Ruusalepp. Assessment of UKDA and TNA compliance with OAIS and METS standards. In *Colchester: UK Data Archive*. 2005. [Online; accessed 31-March-2006].

[31] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. G. Cuesta, J. M. Gomez-Perez, G. Klyne, K. Page, M. Roos, J. E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. A. Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In *ESWC2012 Workshop on Semantic Publication (SePublica2012)*, 2012.

[32] K. Belhajjame, G. Klyne, D. Garijo, O. Corcho, E. García Cuesta, and R. Palma. Wf4ever Research Object Model. http://wf4ever.github.io/ro/, May 2013. [Online; accessed 20-May-2013].

[33] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, MarMar 2009.

[34] L. Candela, D. Castelli, N. Ferro, G. Koutrika, C. Meghini, P. Pagano, S. Ross, D. Soergel, M. Agosti, and M. Dobreva. *The DELOS Digital Library Reference Model Foundations for Digital Libraries*. DELOS, Italy, Dec. 2007.

[35] P. Ciccarese. The Annotation Ontology. http://code.google.com/p/annotation-ontology. [Online; accessed 20-May-2013].

[36] D. De Roure, K. Belhajjame, P. Missier, J. M. Gómez-Pérez, R. Palma, J. E. Ruiz, K. Hettne, M. Roos, G. Klyne, and C. Goble. Towards the preservation of scientific workflows. In *8th International Conference on Preservation of Digital Objects (iPRES 2011)*, NOV 2011.

[37] D. De Roure, C. Goble, and R. Stevens. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, (25):561–567, 2009.

[38] D. De Roure, K. Page, R. Palma, and P. Hołubowicz. D1.3v2 Wf4ever Architecture - Phase II. Technical report, University of Oxford, March 2013.

[39] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, C. Goble, M. Livny, L. Moreau, and J. Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, dec 2007.

[40] R. González-Cabero, R. Palma, and E. García Cuesta. D3.2v1: Design, implementation and deployment of workflow evolution, sharing and collaboration components. Technical report, July 2012.

[41] M. Hedstrom. Digital preservation: A time bomb for digital libraries. *Language Resources and Evaluation*, 31(3):189–202, May 1997.

[42] K. Page, D. De Roure, and K. Martinez. REST and Linked Data: a match made for domain driven development? In *2nd International Workshop on RESTful Design*, March 2011. Event Dates: 28/03/2011.

[43] K. Page, R. Palma, P. Hołubowicz, G. Klyne, S. Soiland-Reyes, D. Cruickshank, R. González-Cabero, E. García-Cuesta, D. De Roure, and J. Zhao. From workflows to Research Objects: an architecture for preserving the semantics of science. In *Proceedings of the 2nd International Workshop on Linked Science*, NOV 2012.

[44] L. Richardson and S. Ruby. *Restful web services*. O'Reilly, first edition, 2007.

[45] R. Ruusalepp and M. Dobreva. Digital Preservation Services: State of the Art Analysis. Technical report, DC-NET, 2012.