DataVerse - SWH SPECS

https://hedgedoc.softwareheritage.org/fc4e-wp6-t6-1-specs-template?edit

Component	*DataVerse software source code & metadata deposit*
Category	***RSAC***
Contact person	*Wim Hugo*
Email address	*wim.hugo@dans.knaw.nl*
Contributors	*Wilko Steinhoff*
Version	*1.0*
Data	

Overview

Repository software will interact with the services created in FAIRCORE4EOSC in multiple ways, and at multiple points during specific workflows. These workflows are described as individual user stories below.

The user stories are specific to repository operations, and not all of the requirements expressed will be satisfied or addressed by FAIRCORE4EOSC components. The user stories are documented as requirements - functional and otherwise - and a high-level architecture/ specifications are provided.

In the detailed requirements and specifications, the elements that can be provided by FAIRCORE4EOSC components and services are identified clearly.

Specifically, some of the services, as indicated, are provided directly by WP6.

Omissions and Work to be Done

1. *Review of the use cases against the work of RDA*
    a. *Publishing Data Workflows*
    b. *Standardised RDA Use Cases*
    c. *Matrix of Use Cases*
    d. *FAIR Workflows*
2. *RDA Repository Terminology Verification (Case Statement)*

Objectives

#	*Description*
1	*Implement workflow steps and interfaces to integrate FC4E services into the main tasks of a repository:* *1.1 Manual and automated deposit of research output* *1.2 Curation - manual and automated* *1.3 Discovery and re-use*
2	*Provide a High-Level Architecture* *2.1 Context - including design considerations* *2.2 Containers and Components -indicating the contribution from FC4E services and components*

Objectives	
3	*Specifications* *3.1 Process flows for major use cases* *3.2 Links to applicable specifications for APIs and interfaces*

Out of Scope	
#	*Short description*
1	*The workflow steps from a repository perspective are reproduced below for sake of completeness, but will not all be in scope for the FC4E project. Those that are not in scope are indicated explicitly*
…	

Requirements

User stories		
#	*Description of the user story*	*Reference*
1	*Manual/ automated deposit of research output* *A user (researcher at a RPO) wants to deposit a collection[1] of research outputs to an appropriate repository for the type of output, and in alignment with institutional and disciplinary practices. The repository should be certified as trustworthy, or, alternatively, provide for some important aspects of trustworthy repositories – such as long-term preservation.* *Ideally, researchers would want to engage the deposit pipeline from within their own working environments: VREs, institutional or disciplinary websites, scientific workflows and code, etc.* *On deposit, the researcher would like to receive guidance in respect of the most appropriate repository for each object in the collection, while providing the minimum required metadata (FAIR, reproducibility[2], schema compliance, PID registration compliance, disciplinary norms, ...) only once across all of the different types of research outputs in the collection.* *In addition, guidance in respect of minimum metadata, appropriate licence selection, and FAIRness of objects in the collection will be very useful.* *The researcher would want to receive confirmation of deposit and associated PIDs at the earliest opportunity, and be able to see an inventory of published works across different repositories.* *1. Typical collection can comprise of one or more of the following:* *a. Input/ raw dataset(s)* *b. Derived/ processed dataset(s)* *c. Code* *d. Semantic artefacts* *e. Reports, papers, and other scholarly publications* *2. Reproducibility requires additional PID references for*	*RDA Use Cases* *RDA Publishing Workflows* *FAIR Workflows* *RDA Repository Attributes* *PresQT*

User stories

	a. Instruments b. Methodology c. Samples, Physical Objects, and Specimens d. Computing Platforms	
2	Curation Tasks Curators require development and integration of new tools to assist with the processing of new and legacy deposits, covering the following broad requirements: 1. Evaluation of FAIR compliance, PID Policy compliance, reproducibility, and alignment with disciplinary norms - ethics, formats, etc. 2. Verification of and linking to controlled vocabularies and registries, and recording PID relations implied or explicitly provided in the metadata. 3. Maintaining provenance and versioning information 4. Updating external semantic artefacts, such as PID and Research Graph services, with information about the deposited object.	Signposting OAI-PMH Linked Data Notifications
3	Enhanced Discovery 3.1 Indexing catalogues with additional facets obtained from external semantic artefacts - primarily but not limited to vocabulary or registry services, and PID/ research graph services. 3.2 Provision of contextual information to end users in respect of repository objects (deposits) - for example compliance or user feedback - from services such as PID and research graphs. 3.3 Embedding graph-based UI components into repository discovery tools to improve utility for the end user.	

User requirements

These are specific user requirements that can be addressed by FAIRCORE4EOSC components or products

#	Short description	Priority	Feasibility	Reference
1	An API for the provision of metadata and references to code for long-term preservation in Software Heritage. The code is linked to a dataset and/ or other research objects being deposited. Needs to return a PID (SWHID) for inclusion into PID graphs and provenance information in Dataverse. PresQT as a candidate specification or facade.	H	4	WP6 Component PresQT
2	An API for retrieving metadata in an agreed schema (CodeMeta) from SoftwareHeritage based on a SWHID, and obtain a citation in one or more standardised formats	H	5	WP6 Component
3	An API for updating PID Graph information based on the LOD references obtained during the deposit process. The source of the LOD contribution needs to be persisted in the target PID Graph.	H	3	

User requirements

4	*An API for updating Research Graph information based on the LOD references obtained during the deposit process. The source of the LOD contribution needs to be persisted in the target PID Graph.*	*H*	*3*	
5	*An API for referencing semantic artefacts in a consistent and sustainable way - can be a brokering platform as envisaged in part for FC4E*	*M*	*2*	
6	*An API for querying and accessing information stored in the Research Graph service developed by FC4E*	*H*	*3*	
7	*An API for querying and accessing information stored in the PID Graph service developed by FC4E*	*H*	*3*	
8	*An API for querying and assessing level of PID compliance of the PIDs included into a metadata record.*	*H*	*3*	
8	*An API for transformation of a metadata record - schematically and/ or semantically - by using the metadata crosswalk service envisaged for FC4E*	*M*	*2*	
9	*An API for registering a concept or type, with a PID returned, in a vocabulary or type registry.*	*M*	*3*	
10	*...*			

- Reference - <JIRA issue number> or <URL to external reference>
- Priority: H=High, M=Medium, L=Low
- Feasibility: marking between 1 - 5, 5 is easy to implement and 1 is very difficult

Functional requirements

<table>
<tr><th><i>#</i></th><th><i>Short description</i></th><th><i>Priority</i></th><th><i>Reference</i></th></tr>
<tr><td><i>1</i></td><td><i>Ingest Workflow</i>
1. <i>Selection of appropriate pipelines, target repositories for collection of research outputs</i>
2. <i>Verification of metadata completeness</i>
3. <i>Assessment of deposited outputs in respect of compliance with multiple sets of criteria</i>
a. <i>FAIR</i>
b. <i>Reproducibility</i>
c. <i>PID Policies</i>
d. <i>Format and schema compliance</i>
e. <i>Optional format conversions for dissemination</i>
f. <i>Optional metadata schema crosswalks</i></td><td><i>[H|M|L]</i></td><td><i><JIRA issue number> or</i>
<i><URL to external reference></i></td></tr>
</table>

Functional requirements

	g. Ethics clearance and protection of subject's rights (CARE) *4. License assessment and evaluation* *a. Rights to allocate or select a license* *b. Extent of private or sensitive data* *c. License selection and applicability*		
2	*Long-Term Preservation Workflow* *1. Creation of preservation metadata* *2. Assessment of deposited outputs in respect of compliance with multiple sets of criteria* *a. Format and schema compliance* *b. Optional format conversions for preservation*		
3	*Curation Workflow - New Deposits*		
4	*Curation Workflow - Legacy Deposits*		
5	*Indexing and Faceting Workflow*		
6	*Contextualisation Workflow*		
7	*Graph-Based Discovery Assistance Workflow* *Repository software such as Dataverse needs to be extended to allow externally provided search and discovery UIs to be embedded into the main application. These extensions invoke the standard Dataverse Search API after assisting the user with query formulation in any of a number of ways.* *1. Select Graph-Based Search Option as an advanced search method in the Dataverse UI.* *2. Navigate a graph of choice (e.g. Research Graph, PID Graph) and formulate a query based on nodes and relations.* *3. Translate the query into corresponding search API facets.* *4. Execute via standard Dataverse Search API.*		*Dataverse Search API*

- Priority: H=High, M=Medium, L=Low

Non-functional requirements

#	*Short description*	*Priority*	*Reference*
1	*<Provide a short description of non-functional requirements>*	*[H|M|L]*	*<JIRA issue number> or* *<URL to external reference>*
…			

- Priority: H=High, M=Medium, L=Low

Specifications

Architectural design
<Provide a high-level architectural design diagram of the component/service>

Functional specifications			
#	Short description	Priority	Reference
1	<Provide a short description of functional technical specification>	[H|M|L]	<JIRA issue number> or <URL to external reference>
…			

- Priority: H=High, M=Medium, L=Low

Service specifications			
#	Short description	Priority	Reference
1	<Provide a short description of service specifications>	[H|M|L]	<JIRA issue number> or <URL to external reference>
…			

- Priority: H=High, M=Medium, L=Low

Operational specifications			
#	Short description	Priority	Reference
1	<Provide a short description of operational specifications>	[H|M|L]	<JIRA issue number> or <URL to external reference>
…			

- Priority: H=High, M=Medium, L=Low

Integration with EOSC Core components			
#	Short description	Priority	Reference

Integration with EOSC Core components			
1	*<Provide a short description of operational specifications>*	*[H|M|L]*	*<JIRA issue number> or* *<URL to external reference>*
…			

- Priority: H=High, M=Medium, L=Low

External references

External references - subcomponent name		
#	*Short description*	*Reference*
1	*<Provide a short description of external reference>*	*<url>*
…		

External references - SWH		
#	*Short description*	*Reference*
1	*<Provide a short description of external reference>*	*<url>*
…		