1 PID Collections WG charter

WG candidate co-chairs: Bridget Almas (Tufts/Perseus DL), Tobias Weigel (DKRZ), Tom Zastrow (RZG)

1 Value Proposition

Several communities have expressed a need to leverage aggregations of objects with a particular focus on building such aggregations, whether virtual or physical, through PIDs and providing identifiers for aggregation objects. There is however no unified cross-community approach to building and managing such collections and no common model for understanding them. The PID Information Types WG has defined a core model and the central interface for accessing object state information and provided a small number of example types, which were consequently registered in the Data Type Registry WG prototype. With these tools available to describe essential object information, collections can be described so to be able to deal with more than a single object at once.

Building collections within diverse domains and then sharing or expanding them across disciplines should enable common tools for end-users and e-infrastructure providers. Individual disciplinary communities can directly benefit if such tools are made widely available, and cross-community data sharing can benefit from increased unification between collection models and implementations. PID providers may benefit from marketing additional services on collections.

2 Engagement With Existing Work

3 The WG will examine existing models for identifying and managing collections to surface commonalities and differences across models and to ensure that the output of the WG is general enough to work with these standards. Specific standards that we will investigate include the IETF BagIT Draft specification[1], the CITE Collection Services protocol[2] and OAI-ORE[3]. It is not the intent of the working group to propose an alternative to existing well established standards for describing and archiving collections but rather to propose an API and implementation for creation, consumption, distribution and citation of collections and their items that could serve as a unifying layer on top of the existing models.

The WG will observe other developments within and outside of RDA such as the ongoing Type Registry work and similar typing efforts. The later phases of the WG effort may also coincide with concerns within the EUDAT2 project. The notion of collections has also been included in the first model discussions of the Data Fabric IG, and the WG will contribute to these discussions.

4 Goals and work plan

The WG will start with an assessment of community use cases, some first examples are given further below. From the use cases, a classification scheme or general model should be developed that explains the different approaches and understandings in describing collections, including aspects such as static and dynamic collections. Another important model to recognize during WG work are collections based on file system directories as these represent today's most common approach to

1 https://tools.ietf.org/html/draft-kunze-bagit-10

2 http://cite-architecture.github.io/cc_spec/

3 http://www.openarchives.org/ore/

organizing data. Eventually, such models may contribute to a view where digital objects and collections become the equivalent to traditional files and directories.

For a choice of the use cases, the respective collection models should be expressed through PID types and these types should be registered. Other relevant candidate types that go beyond core collection concerns may be discussed as well. Discussions may also cover other methods to relate objects to each other in general object or identifier graphs, building on prior work e.g. in the context of RDF/OWL or FRBR. As part of this discussion, the role of identifier fragments and queries in the collection models should be clarified, and models for fragment services should be discussed. The selected use cases then feed into the formulation of a generic collection API, extending and unifying existing solutions (e.g. from CLARIN or OAI-ORE). Possible themes for the API also include methods to differentiate between nodes and leafs, supported by specialized PID types, and to offer iteration and traversal operations. With respect to such a unifying API and the community use cases, added-value tools should be discussed that offer direct benefits to community end-users. The collection API should be implemented in a small demonstrator project which may also illustrate some tool ideas. To work across identifier systems, the demonstrator should make extensive use of the PID Information Types API. The most essential typing mechanisms that can be used to implement collections should be registered in a Type Registry.

The WG aims to have a productive working session at each of the corresponding RDA plenaries. Besides members from infrastructures and PID providers, representatives from user communities are particularly welcome. Between plenaries, WG work will continue in small groups via e-mail and virtual meetings.

5 Expected concrete outcomes

D1. Collection models (M12). This report summarizes the collection models with detailed descriptions and usage examples and should help communities to understand and refine their collection usage scenarios. Fragment identifier issues will be addressed as well. This should be a step-by-step guide to the what, why and how of collections.

D2. API and demonstrator (M18). This deliverable includes the collection API specification, documentation and a demonstrator that illustrates the added value of unified collections. A final list of suggested PID types should be included. Paper prototypes for tools or other applications within exemplary domain scenarios may also be provided. Although development of the API specification will only be done through detailed analysis of the use cases, we envision that at a minimum the following types of collection operations would be covered:

- Retrieving/setting/updating collection level metadata
- Retrieving a list of items (ordered or unordered) in a collection
    - refinements on this will include pagination and filtering by specific criteria
- Create/Read/Update/Delete operations on collection items

We expect some more advanced requirements to be uncovered as well through the use case analysis, such as capabilities for discovery of fragment identifiers and definition of collection type templates.

In accordance with the guidelines of RDA, all outcomes will be provided under open licenses.

1 Social Deliverables and Sustainability

As described above, the working group plans to deliver an easily adoptable model for identifying and managing collections of data objects via the combination of a clear outline of use case scenarios, a well-defined API for machine driven interaction with the collections, and a reference implementation of that API deployed by projects across several domains. We hope that this can provide a straightforward solution for many research projects that might otherwise have implemented a closed or idiosyncratic model for their data collections. We expect that this will be a living solution, which is improved over time by the addition of new use cases to make it more robust. A focus over the WG lifetime is to keep the rather abstract API design and the concrete domain use cases closely connected, e.g. by including textual usage scenario descriptions from the individual users' point of view that show exactly which parts of the API are used (and how they are used) in an exemplary realistic workflow.

6 Milestones

M1: BOF at RDA Plenary P6, additional adopters identified and committed

M6: Initial use case descriptions gathered.

M12: Collection models defined. Collection API draft reviewed.

M18: Collection API and demonstrator implemented.

7 Adoption Plan

The following organizations have expressed a commitment to adopt the outputs of the WG, by implementing and deploying the API for their specific use cases:

- **DKRZ:** Collections are useful to bind dataset replicas and versions together and reflect the multi-hierarchical organizational structure of the ESGF dataspace. Such collections are largely static, but highly interconnected with other collections and objects. Implementation continues throughout 2015 and some essential collection tools may be developed in the time afterwards.

- **Perseids Project, Perseus Digital Library:** The Perseids Project at the Perseus Digital Library requires an application of collections and fragments for referencing (human & machine), not for object management (moving objects around in an e-infrastructure). Collections are built explicitly via hierarchical PID syntax components which are widely agreeable to be static; for machines, a common API would unambiguously expose the hierarchical levels. Annotation types could be expressed through PITs.

Additionally, the following organizations have expressed a commitment to supplying use cases for the WG and to strongly consider adoption of the outputs:

- **BCO-DMO:** The use case focuses on cruise data acquired with various instruments used also across several cruises. Users may perform diverse discovery and aggregation tasks, e.g. for data from a single cruise or the same instrument used across several cruises. Data objects are accordingly arranged in collections; sometimes hierarchical but more often graph-like depending on use. WHOI is looking into assigning specific PIDs to cruise data (DOIs) and

related concepts (e.g. ORCIDs for person, IGSNs for physical samples) and interconnecting them.

- **SEAD project:** The SEAD virtual archive offers several relevant workflows for research object management, including collection and subcollection building and versioning. A demonstrated practical use case showed the importance of building a virtual collection with data objects of mixed type. From the view of the Collection WG, a first opportunity is the common collection API.

- **Coptic SCRIPTORIUM:** The Coptic SCRIPTORIUM project has provided a use case centered on identifying, referencing and managing collections of textual and linguistic data objects, including codices, paleographic symbols, manuscript fragments, digital images, annotations, morphemes and word tokens. These data objects are currently managed by the project through spreadsheets without use of persistent identifiers and a scalable solution is required to manage this data and it to the PIDs of the source texts.

- **Ocean Data Interoperability Platform (ODIP).** For the purposes of data publishing, the ODIP project is creating data set collections using pre-defined criteria such as vocabulary terms or originating repositories. Usually, both collections and their granules bear PIDs.

- **Harvard Astronomy Abstract Service:** Tables with data values as supplements to articles, and individual values in table rows, articles bearing PIDs. PID fragments should then point to individual values, which enables better data discovery. The practical feasibility depends on the uniformity of table encodings.

- **Open Philology Project, University of Leipzig:** The Open Philology project has collections of various types of data objects relating to texts (e.g. manuscript images, OCR output, TEI XML, and annotations). They want to be able to apply persistent identifiers to these collections and their objects, as well as to the primary sources to which they refer, throughout the data production, publication and preservation lifecycle. Annotations and derivative versions and analyses which are made on the early versions should be easily and automatically portable to the newer, improved versions as they become available. Citations which reference fragments of the text should be robust and automatically resolvable across versions and archived copies.

  Other interested parties:

  EUDAT, DARIAH and CLARIN have relevant use cases and have expressed interest in following the activities of the WG.

8 Initial Membership

Working Group Co-Chairs,

Tobias Ziegel, DKRZ (German Climate Computing Center)

Thomas Zastrow, RZG Max Planck Society

Bridget Almas, Tufts University, Perseus Digital Library, Perseids Project

Use Case Providers, Potential Adopters:

Beth Plale, Indiana University, SEAD Project

Cynthia Hudson Vitale, Washington University St. Louis

Cyndy Chandler, Biological and Chemical Oceanography Data Management Office, Woods Hole Oceanographic Institution (BCO-DMO/WHOI)

Helen Glaves, British Geological Survey. Ocean Data Interoperability Platform project (ODIP)

Caroline T. Schroeder, University of the Pacific, Coptic SCRIPTORIUM Project

Giuseppe Celano, University of Leipzig, Open Philology Project