# Research Data Collections WG charter

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

## 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 (PIT) 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 (DTR) prototype. With these tools available to describe essential object information, collections can be described so as to be able to deal with more than a single object at once. The new Working Group will therefore also contribute to the evaluation of the outputs of these two prior groups.

As described further below, and as evidenced by the use case details, we believe the outputs of this Working Group can directly lead to the following outcomes:

* Better data management practices result in data that can be more easily shared.
* The ability to add meaning to data by grouping into collections, virtual or otherwise, encourages data sharing and reuse by allowing it to be used for means other than that intended by the original creator.
* Clear definition of collection models, and a common CRUD API for data collections allows for quicker incorporation of new, possibly anticipated data collections into research tooling support, thereby increasing the speed and scale at which data which can be shared and reused.
* A multi-disciplinary approach to collection management results in greater sustainability of the supporting software.

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. Including set, filter and mapping operations in the Collections API would provide a common platform for curation through mixing and mapping of previously defined collections. PID providers may benefit from marketing additional services on collections.

One concrete example for a benefit that exceeds the capabilities of a single community can be found in the Earth System Grid Federation, a global climate data e-infrastructure. The challenges are generic enough not to be limited to this particular community. In the concrete case, individual data objects are aggregated into multiple hierarchical layers, each of which can be individually addressed and used by data consumers. Some of these aggregations remain static, once created, while others grow irregularly over time due to data revisions published in the infrastructure. There are numerous concrete use cases from the humanities where there is a need for a common interface to multiple different models of collections, including virtual, and collections of like as well as diverse data items. Collection data members may also belong to other collections, often leading to mixed models for access. While the developer community can invent the necessary schemas and supporting tools to manage these collections, their application will be limited since many users work with collections and their elements outside the scope of the primary infrastructure. At the same time, maintaining tools for collection management is a costly effort that can easily exceed the capabilities of an individual institution or community. A more general approach can thus evoke long-term cost benefits and gains in usability if collection management and end-user tools interoperate across separate community infrastructures.

## Engagement With Existing Work

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 W3C Linked Data Platform (LDP) recommendation[[1]](#footnote-1), the IETF BagIT Draft specification[[2]](#footnote-2), the CITE Collection Services protocol[[3]](#footnote-3) and OAI-ORE[[4]](#footnote-4). We will also engage with the DDI modeling group that has done some work on modeling collections[[5]](#footnote-5). 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. Existing solutions focus on describing collections and their semantics with metadata, but do not offer a full set of generic CRUD operations on them, which is a key innovation of the proposed API.

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. We will also connect with the proposed Metadata Standards Catalog WG and recently completed Data Description Registries Interoperability WG for opportunities for cross-group collaboration.

## 3. Goals and Work Plan

The WG will start with an assessment of community use cases, some first examples are given further below. Our BOF discussions at Plenary 4 have helped us refine the questions we need to ask of the community in assessing the commonalities and differences across use cases and prioritizing requirements. We understand in particular that we (1) need to clarify the demarcation point between collection, PID and data object[[6]](#footnote-6), and the division of responsibility between a collections API and the PID system, (2) need to identify the initial target audience and determine whether we must prioritize general usability over highly specific technical uses. We have issued a survey to the participants of our BOF asking the following key questions:

* Will your objects bear PIDs?
* Will your collections bear PIDs?
* Do you need a collections solution for your internal object management?
* Do you need a collections solution for interoperability with external partners?
* Do you already plan to implement a collections solution within the next 24 months and how concrete are these plans?
* What kinds of objects do you see in your collections?
* Can you imagine concrete examples for your use case for end-user tools or services that build on an outcome of this WG?

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 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 leaves, 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.

## 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. This should be a step-by-step guide to the what, why and how of collections. The models proposed as part of the PIT and DTR group deliverables will be reused; the deliverable may propose extensions if required.

**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.

## 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.

## 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.

## 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:** Collectionsare useful tobind 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. (For more on the Perseids use case see http://bit.ly/1Q3rBMZ).
* **NoMaD Repository:** The NoMaD (Novel Materials Discovery) Repository was established to host, organize, and share materials data. The atomic datastructure in the NoMaD Repository is a "calculation". For achieving a higher degree of organization, these calculations can be grouped into "datasets". A dataset can contain other datasets which allows creation of a recursive tree-like structure of datasets. Datasets can be used to filter the results of a (faceted) search. Datasets in the NoMaD Repository can be linked to existing publications via their DOI. Optionally, a new DOI can be assigned to a dataset.
* **PECE Project, Rensselear Polytechnic Institute:** Needs to manage “bundles” of data object “artifacts” requiring very fine-grained levels of attribution and identification. These bundles are collections of heterogenous objects, including images, audio, video, text, annotations, web pages and other collections. The value proposition of a collection is that it adds meaning to data. The ability to make collections out of items diverse other collections would support data sharing because it allows researchers to use data in ways other than originally envisioned by the creator. For this use case it is necessary to support creation of collection of items which themselves have not yet been persistently identified, making collections a tool for data creation.

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

* **ORCID, Inc.** ORCID records can themselves be thought of as collections of data about an individual researcher. They aggregate heterogeneous data objects including biographical data, education history, research activities, funding sources, works produced, peer reviews of which themselves may be represented as collections. These collections have complex access rules that need to be enforced for both create and read operations.
* **CLARIN**: CLARIN currently supports the concept of “Virtual Collections” which are sets of links to digital objects that can originate from different archives. CLARIN provides a registry where scholars can create and publish their collections, and which provides PIDs and federated access. Development and adoption of a common collections API would facilitate sharing of collections managed by CLARIN centers with other repositories and domains.
* **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:** 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.
* **IGSN:** has a need to manage and share collections of Geo Samples which are identified by persistent identifiers in the form of ISGN numbers.
* **Natural History Museum London, Department of Life Sciences:** is currently going through a re-evaluation of workflows and tools for collection management and would like to be able to include the outputs of this WG in the solutions they consider.

## Initial Membership

*Working Group Co-Chairs,*

Tobias Weigel, DKRZ (German Climate Computing Center)

Thomas Zastrow, RZG Max Planck Society

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

Frederik Baumgardt, Tufts University, Perseus Digital Library, Perseids Project

*Use Case Providers, Potential Adopters:*

Kim Fortun, Rensselear Polytechnic Institute, PECE Project

Beth Plale, Indiana University, SEAD Project

Dieter Van Uytvanck, CLARIN

Laura Paglione, ORCID, Inc.

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

Dimitris Koureas, Natural History Museum London

Wendy Thomas, Data Documentation Initiative (potential – has expressed interest in participating)

1. http://www.w3.org/TR/ldp/

   [↑](#footnote-ref-1)
2. 2  https://tools.ietf.org/html/draft-kunze-bagit-10 [↑](#footnote-ref-2)
3. 3 http://cite-architecture.github.io/cc\_spec/ [↑](#footnote-ref-3)
4. 4. http://www.openarchives.org/ore/ [↑](#footnote-ref-4)
5. https://ddi-alliance.atlassian.net/wiki/display/DDI4/Modelling+Team [↑](#footnote-ref-5)
6. i.e. where/how to draw the line between what is metadata on the PID and what is the property of the data object itself and what is the property of the collection? A particular point of concern may be data access and authentication/authorization, for which the API may provide a dedicated metadata section that external AAI services can use. [↑](#footnote-ref-6)