Case Statement/Charter
for the establishment of an
Joint RDA/TDWG Working Group
on

Metadata Standards for attribution of physical and digital collections stewardship

Thessen, A.E., Woodburn M., Ariño A., Flann C., Nicolson N., Shorthouse D. and Koureas D.

1. WG Charter

This working group (WG) will address the incomplete standards for giving attribution for the maintenance, curation, and digitization of collections. Within the scope of the WG, collections can include digital data and digital/physical objects. This WG will produce use cases from a variety of disciplines that will be used to create the final deliverable – an attribution metadata schema. Adopters will include stewards of collections (such as the Natural History Museum London) and aggregators of professional research metrics (such as ImpactStory).

2. Value Proposition

Research collections are an important tool for understanding the Earth, its systems, and human interaction. These collections are very diverse and can include preserved natural history specimens, archeological artifacts, or historical documents, to name just a few. Maintaining and curating these collections requires a large investment of time and money by institutions and many individuals. Knowledge is created from collections by many individuals over time, building on the work of others. For maximum efficiency, work needs to be shared broadly, recorded permanently, and tasks not repeated unnecessarily. Unfortunately, the current research cyberinfrastructure does not support this level of efficiency.

Despite the importance of collections, many are not maintained or curated as thoroughly as they should. Part of the reason for this is the lack of professional reward for curatorial actions. Most of the researchers who are qualified to curate a collection are too busy performing activities that will reap professional reward, such as publication and grant-writing. Proper methods of attribution (at the individual and institutional level) are very important for incentivizing digitization, mobilization. And sharing of data deriving from collections (physical and digital). One strategy for incentivizing physical and digital collection curation is to create infrastructure for attributing curatorial actions. Several programs exist for aggregating metrics for research products other than publication, such as ImpactStory, OpenVIVO, Collector, and Altmetrics.

Thus, there is already infrastructure in place for aggregating these data, if the e-infrastructure for creation of these data is available.

Significant investment has been made in creating infrastructure components for data integration across a wide variety of disciplines. Many of these components are lists, repositories, or other structures that must be populated with data either by a person or algorithmically. Even an automatically-created data set will require some degree of human curation to ensure quality. Often, very little can be completed without initial work by a person to create reference material. This human-component is a major bottleneck. Thus, existing infrastructure for collective resources are not being populated with data and thus are not maximally useful. One way to widen the bottleneck is to create professional incentives for researchers to contribute to maintaining and curating collections. If people could get professional credit for improving a classification, for example, it would be much easier for them to dedicate the time required. The problem is that there is no good way to manage information about curatorial actions so that curators can get professional credit.

The goal of this WG is to develop an attribution data schema (in collaboration with adopters and with use cases from several disciplines) that can make getting credit for curation, maintenance, and digitization of a collection as easy as getting credit for a publication. The deliverables of this WG will benefit institutions that maintain collections and individuals who curate them and will lead to:

- Improved recognition of the immense effort required for maintaining, curating, and sharing collections, which is likely to lead to increased funds for these activities
- Increased efficiency in knowledge generation from collections through the proper documentation of corrections and analyses performed
- Increased viability of crowdsourcing as a model for building collaborative research resources
- Increased relevance of existing e-infrastructure that is being stifled by the expert annotation bottleneck

3. Engagement with existing work in the area

Most institutions that maintain collections of physical items employ, or are moving towards, a central Collections Management System (CMS) to support digital object curation. Certain information about personal contribution to digital activities can often be assembled from the generic database audit trail incorporated into these systems. However, the primary function of these structures is to support system and workflow requirements, so are rarely able to provide complete and accurate attribution metadata, and can only reflect digital rather than physical effort. There is therefore a strong case for an attribution metadata schema which these systems

could adopt as part of their data model and workflows. Several vocabularies have been developed specifically for recording provenance information (PROVO), for recording information about physical samples (IGSN), and for describing contributor roles (TaDiRAH, CRediT, OpenRIF). In addition, several domain-specific standards provide methods for giving attribution for a physical object, data set, or data product (TDWG, ESIP, SESAR, CODATA, COPDESS, etc.). None of these standards provides a method for recording specific curatorial actions on a physical/digital object, digital data set, or data product. All of these existing standards provide pieces of a system that, with some additional work, could make attribution and professional reward for curatorial actions possible. This WG will strive to ensure interoperability between its recommendations and existing schema.

The PID (Persistent Identifier) Collections WG, which currently has its case statement in review, is potentially very relevant to the work we propose. Briefly, this group will develop collections-level metadata and specifications for an API. This WG will be more focused on developing metadata for individual objects within a collection rather than collection-level metadata, but we will collaborate with this group to ensure that our schema are interoperable. The Metadata Standards WG and Interest Group (IG) are very relevant to the goals of this proposed WG. We will provide a use case for these groups and align our schema with the proposed metadata elements.

One group with whom we will be working very closely is IGSN, an organization that provides a unique identifier for physical samples. This group started working primarily with geological samples, but are now moving toward accommodating biological samples. We feel that their initial metadata schema is a good starting point.

Museums, repositories, and other stewards of collections are always working hard to maintain and curate their collections for maximum use. This WG will be pursuing these institutions as adopters and working closely with them to investigate large-scale viability of solutions they have implemented as well as ensuring WG deliverables will be useful to them.

In order to have a true impact on the social aspect of professional reward, the WG deliverables need to ensure that data within the schema can be used by professional metrics aggregators such as ImpactStory. We will work closely with this project to make sure that their system can handle our products. One important difference between this WG and other efforts is the focus on outputs that result in actionable metrics.

4. Work plan

The work of this WG will be completed in 18 months. We will split the tasks and milestones into three concurrent work packages (WP).

Work package 1: Requirements (M1-M6)

- Task 1.1. Develop use cases via WG contribution and community engagement
  Milestone 1.1: Use cases report (M6)

Work package 2: Technical (M4-M16)

- Task 2.1. Investigate existing schemas/infrastructure
  Milestone 2.1: List of relevant references (M6)

- Task 2.2. Develop attribution metadata standard and schema
  Milestone 2.2.1. Draft attribution metadata standard and schema document (M12)
  Milestone 2.2.2. Schema review with feedback from case studies (M16)

Work package 3: Community establishment (M6-M18)

- Task 3.2. Initiate process suggesting the schema for ratification as a community standard through TDWG
  Milestone 3.2. Process initiated, and acknowledged by TDWG (M17)

- Task 3.1. Liaise with stakeholders and community actors to establish adoption plans through piloting actions
  Milestone 3.1. Report potential adopters and planned actions (M18)

WG final deliverable

- Final attribution metadata standard and schema document (M18)

5. Adoption plan

In order to ensure the eventual functionality and to maximize usefulness of the schema the WG will consult with stewards of physical specimens and aggregators of professional metrics. These collaborations will start from the very beginning of the WG. Potential adopters are categorised in the following groups:

1. Data providers/Data stewards
2. Aggregators/Repositories
3. Publishers
4. Scholarly metrics providers

The following organizations have from the outset expressed interest in the deliverables of the WG:

- Natural History Museum London, UK (Data provider)
- Royal Botanical Gardens Kew, UK (Data provider)
- Biodiversity Heritage Library (Data provider)
- Pensoft (Publisher)
- ImpactStory (metrics provider)
- Naturalis (Data provider)

6. Operational Policies

6.1. WG mode and frequency of operation

This WG will hold in-person meetings at RDA plenaries as well as TDWG meetings. Also virtual meetings will be held every month. Virtual meetings will be recorded and posted for interested parties who could not attend. Every three months a short report on activity will be requested by the WP leaders and circulated to all members of the WG. All WPs will be supported through a wiki, a developer forum, and mailing lists.

6.2. Plans to develop consensus, address conflicts, and stay on track

All meetings will be kept on track by having an agenda, action items, and deadlines for those action items. The deadlines will not be flexible. In the event that there is still a lot of open discussion as a deadline approaches, the state of discussion will be reported in the corresponding deliverable. Consensus will be reached via open discussion and voting as appropriate. It is the responsibility of the WG leaders to build consensus through structured moderation. If a conflict cannot be resolved within the WG, the RDA council will be consulted and an independent party will be brought in to mediate. The WG will avoid mission creep by sticking to the project plan as outlined above. Appointed moderators and WG leaders will enforce focused discussion by, for example, splitting forum threads as appropriate

6.3. Broader community engagement and participation plan

This WG will hold working meetings and joint meetings at every RDA plenary. The monthly meetings will be open to any interested party regardless of WG membership. Notes, slides, and recorded meetings will be made available on the RDA website. The wikis and forums will be open.

7. Initial membership

	Name	Affiliation	Country	role

1	Agosti D.	Plazi	CH	
2	Ariño A.H.			
3	Flann C.	Species 2000	NL	
4	Koureas D.N.	Natural History Museum London	UK	
5	Miller C.			
6	Nicolson N.			
7	Penev L.	Pensoft	BG	
8	Piwowar, H.	ImpactStory	US	
9	Priem, J	ImpactStory	US	
10	Pyle R.			
11	Schentz H.			
12	Shorthouse, D.	Université de Montréal	CA	
13	Thessen A.E.	Ronin Institute for Independent Scholarship and The Data Detektiv	US	
14	Patterson D.J.	Plazi	AU	
15	Woodburn M.S.	Natural History Museum London	UK	
16	Kersten Lehnert	Columbia University	US	
17	Wouter Addink	Naturalis	NL	
18	Stacy Konkiel	Altmetrics	US	

Acronyms
TDWG - Biodiversity Information Standards
ESIP - Earth Science Information Partners
IGSN - International Geo Sample Number
CODATA - Committee on Data for Science and Technology
COPDESS - Committee on Publishing Data in Earth and Space Science

SESAR - System for Earth Sample Registration
RDA - Research Data Alliance