Islam, S, et al. 2020. Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. *Data Science Journal*, 19: 50, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2020-050

**PRACTICE PAPER**

# Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections

Sharif Islam[1], Alex Hardisty[2], Wouter Addink[1], Claus Weiland[3] and Falko Glöckler[4]

[1] Naturalis Biodiversity Center, Leiden, NL

[2] Cardiff University, Cardiff, UK

[3] Senckenberg Biodiversity and Climate Research Centre, Frankfurt, DE

[4] Museum of Natural History, Berlin, DE

Corresponding author: Sharif Islam (sharif.islam@naturalis.nl)

To support future research based on natural sciences collection data, DiSSCo (Distributed System of Scientific Collections) – the European Research Infrastructure for Natural Science Collections – adopts Digital Object Architecture as the basis for its planned data infrastructure. Using the outputs of one Research Data Alliance (RDA) interest group (IG) and five working groups (WGs) we show how RDA recommendations and supporting documents have been applied to the various stages of the DiSSCo data lifecycle.

**Keywords:** Distributed System of Scientific Collections; DiSSCo; Research Data Alliance; RDA; Digital Object Architecture; Digital Specimen

## Introduction

In this paper, we describe how the outputs of one Research Data Alliance (RDA) interest group (IG) and five working groups (WG) have shaped the core concepts of DiSSCo (Distributed System of Scientific Collections)[1] – the European research infrastructure for Natural Science Collections. Designing, building and operating a research infrastructure like DiSSCo, which has a high dependence on information and communication technologies (ICT) and data management best practices brings together expertise from multiple domains (museum curators, taxonomists and other scientists, biodiversity informaticians and data managers, computing and software engineers, administrative management). The complex design decisions involve interrelated technical components spanning five data lifecycle phases, from data acquisition through data curation, data publishing and data processing to data use (Martin et al., 2017; Nieva de la Hidalga et al., 2020: 66–67). The collective expertise from RDA and the published recommendations provides the DiSSCo community with useful guidance for creating and supporting a sustainable, long-lived research infrastructure that can enhance the overall capacity of the user to find, retrieve, and use relevant information. How this community has used RDA recommendations to shape the DiSSCo approach is generic enough to be of interest to readers from other fields.

The paper is organized as follows. We begin with the background on Natural Science Collections (NSCs) in the context of recent advances in digitization, data sharing and how the new challenges in the future can be addressed by a research infrastructure such as DiSSCo. The background also introduces the Digital Specimen concept– a particular type of FAIR Digital Object and the DiSSCo data lifecycle. Foregrounding the DiSSCo data lifecycle then we describe how selected outputs of RDA are applied in the design of DiSSCo data infrastructure. We conclude the paper with an overview of the future core DiSSCo services that these design decisions will enable.

[1] Distributed System of Scientific Collections (DiSSCo): https://dissco.eu.

The RDA outputs cover the following aspects:

1. RDA output dealing with the adoption of Digital Object Architecture, based on the work of the Data Fabric IG and the Data Foundation and Terminology WG (RDA DF&T 2015);
2. RDA output dealing with the usage of persistent identifiers and kernel information in the context of machine actionable services and programmatic decisions for digital objects from the PID Kernel WG (RDA PID KI 2019);
3. RDA output dealing with the aggregation of digital objects in the context of meaningful entities and serving the data from the Research Data Collections WG (RDA Research Data Collections 2017);
4. RDA output dealing with curation and maintenance of digital objects from the RDA/TDWG Metadata attribution WG (RDA/TDWG Attribution Metadata 2018); and
5. RDA output covering guidelines and specifications to assess the DiSSCo FAIR implementation plan (RDA FAIR Data Maturity Model 2020).

## Background

Natural Science Collections (NSCs) hosted in natural history museums, botanic gardens, universities and other research centres around the world contain data that are critical for many scientific endeavours (Hedrick et al. 2020). Over the years various large scale digitization projects (Blagoderov et al. 2012), mobilization of biodiversity data (Nelson and Ellis 2019) and use of museum specimens to study genetic diversity (Nachman 2013) provided novel ways of doing science (Schindel and Cook 2018). Within the context of COVID-19 pathogen discovery research, Cook et al. (2020 p. 2) highlight the crucial role of the information system related to collections that hold specimens:

> *"In the past few decades, museums have become hubs of biodiversity informatics, serving as the critical nexus between biological samples and sample-derived data (e.g., genomics, geographic information, isotope chemistry, CT scans). The current pandemic reminds us that natural history specimens are important but underappreciated reservoirs for studying the hosts and distributions of animal and human pathogens (see Harmon et al. 2019) and that the data connected to these specimens increase our understanding not only of the host organism but of the pathogens as well. Enhanced support of both physical and cyberinfrastructure for biodiversity collections would yield an information system to enable prediction and mitigation of future outbreaks and pandemics."*

To support data infrastructures for collections-based research in the future (and this includes their initial design and implementation), we need to understand the challenges and urgency ushered by the new types of data collection, curation, and sharing (e.g., Kays et al. 2020; Kays, McShea and Wikelski 2020) along with maintaining and providing access to historical data (e.g., Besnard et al. 2016; Lister et al. 2011). The physical materials (samples and specimens stored in natural history museums, seed banks, cryo banks, etc.) are crucial elements for scientific inquiry. However, accessing these physically comes with its challenges of reuse as materials can deplete and the distribution of traits and phenotypes in species populations in living collections varies over time (Diaz et al. 2016). Therefore, access to digitized data acts as an essential reference point to the relationship between the digital and physical world. This anchoring of the different kinds of data derived from physical specimens has been explored and described as the notion of the Extended Specimen (Webster 2017). It represents the integrative and interdisciplinary next generation of NSCs (Schindel and Cook 2018). We use the term 'Digital Specimen' (explained below) in an analogous manner.[2]

Existing systems for exploiting material stored in NSCs are inefficient and not cost-effective (Smith et al. 2019). Despite significant work by global data infrastructures such as the Global Biodiversity Information Facility (GBIF),[3] Biodiversity Heritage Library,[4] and Plazi TreatmentBank,[5] there remain systematic gaps in linking specimen data to other data classes such as DNA sequences, literature, functional traits, habitat and

[2] At the TDWG 2020 conference (https://www.tdwg.org/conferences/2020/working-sessions/#bof01) the similarities and differences between Digital Specimen and Extended Specimen concepts were explored and plans for further global collaboration towards a global standard were discussed.

[3] https://www.gbif.org/.

[4] https://www.biodiversitylibrary.org/.

[5] http://plazi.org/.

conservation data and ecological models (Page 2016; Senderov et al. 2018). We are noticing increased use of digitized data from NSCs (Blagoderov et al. 2012). However, at the same time, for many projects, these data are organized and managed in a manner that makes data linking, sharing, and future reuse problematic (Lewis et al. 2018).

Over the past several years, the community around NSCs recognized the gaps in our understanding of bio- and geo-diversity due to loosely coordinated data infrastructures (Hardisty and Roberts 2013). This has led to increased efforts towards creating shared global roadmaps for biodiversity informatics (Hobern et al. 2019), developing standards for improved data quality (Chapman et al. 2020), adopting FAIR principles (Agosti et al. 2019; Grobe et al. 2019) and creating building blocks for a data landscape in which component systems can exchange and understand the information in a standard form using open protocols, and metadata (Lannom et al. 2020). The Distributed System of Scientific Collections (DiSSCo), along with several global partners, is working towards such a data landscape by building a pan-European Research Infrastructure (RI) that aims to mobilize and unify bio- and geo-diversity information connected to the specimens held in natural science collections. As of February 2020, DiSSCo entered the preparation phase where key design decisions and best practices are influenced by five selected outputs from the Research Data Alliance (RDA) (summarised in **Table 1**) along with the FAIR data principles (Findability, Accessibility, Interoperability, and Reusability) (Wilkinson et al. 2016; Mons et al. 2017) and the concept of FAIR Digital Objects (FAIR-DO) (De Smedt et al. 2020; Wittenburg and Strawn 2019; European Commission 2018).

DiSSCo's vision is to transform a landscape of disconnected individual natural science collection providers into a coherent research infrastructure with a variety of e-services to enable this: 1) the European Loans and Visits System (ELViS),[6] a one-stop shop for access to the collections, providing both physical access and virtual access by digitization on demand; 2) European Curation and Annotation System (ECAS) for community curation of the digitized specimen data; 3) Specimen Data Refinery (SDR) providing digitization services to extract, enhance and annotate data from specimens digital images; 4) Collections Monitoring Dashboards (CMD) showing the digitization status and usage of the collections and 5) a knowledge base providing protocols, digitization resources, manuals and other documents as FAIR-DO for direct integration with the other e-services. The RDA outputs mentioned here are providing essential building blocks for envisioning these services.

**Table 1:** RDA outputs applied to the management of the DiSSCo data lifecycle.

| RDA output | RDA IG/WG | DiSSCo Element | Purpose | Workflow/Data phase |
|---|---|---|---|---|
| 1. Adoption of Digital Object Architecture | Data Foundation and Terminology WG Data Fabric and Terminology IG | DiSSCo Digital Specimen Architecture | Define the FAIR Digital Object Architecture of DiSSCo, including the Digital Specimen Object Model | Creation and management of digital objects/ All phases of the data life cycle |
| 2. Persistent Identifiers and Kernel Information | PID Kernel WG | Meta-information about a digital object and DiSSCo (data) type registry | Allowing smart programmatic decisions and inspection of the object's PID record | Data acquisition, curation, publishing, use |
| 3. Aggregation of digital objects | Research Data Collection WG | DiSSCo data repository/portal/API | Provide meaningful entities and serving the data | Data publishing and use (share, download) |
| 4. Metadata attribution and use of PROV entities | RDA/TDWG Metadata attribution working group | Digital Specimen and collection objects | Correctly attribute sources of data and work carried out | Digitization, curation and maintenance of digital object (for example collection objects or specimens) |
| 5. FAIR data maturity model | RDA FAIR data maturity model working group | DiSSCo Digital Specimen Architecture | Develop guidelines and specifications to assess FAIR implementation plan. | DiSSCo data lifecycle |

[6] Current ELViS development activities can be found in GitHub: https://github.com/DiSSCo/ELViS/wiki.

One of the critical elements in DiSSCo is the 'Digital Specimen', a FAIR-DO acting as a digital twin on the Internet for a specific physical specimen in a museum collection. The digital information derived from the specimens will enable FAIR data and services where various data classes can be linked to provide seamless unified access to information. These ideas were explored in the EU-funded ICEDIG[7] project (2018–2020) and one of the core architecture outcomes was the decision to adopt FAIR Digital Objects (FAIR-DO) (Hardisty et al. 2020). In particular, this choice enables the creation of machine-actionable digital twins which by design ensures FAIRness[8] of the data and various other features such as unambiguous identification, data typing enforcement, attribution and provenance tracking.

The following sections explain how five selected RDA outputs, summarised in **Table 1**, are applied to the management of the data lifecycle in DiSSCo, illustrated in **Figures 1** and **2**.

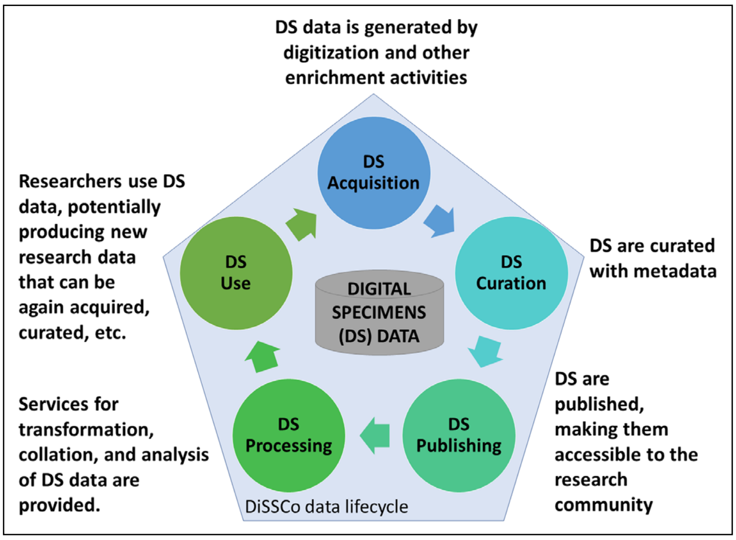
DS data is generated by digitization and other enrichment activities
DS Acquisition
DS Curation
DS are curated with metadata
DIGITAL SPECIMENS (DS) DATA
DS Publishing
DS are published, making them accessible to the research community
DS Processing
Services for transformation, collation, and analysis of DS data are provided.
DS Use
Researchers use DS data, potentially producing new research data that can be again acquired, curated, etc.
DiSSCo data lifecycle

**Figure 1:** Lifecycle of Digital Specimen research data in the DiSSCo data infrastructure, from acquisition through curation, publishing, processing and use, which can create new data that can be iteratively acquired, curated, etc.

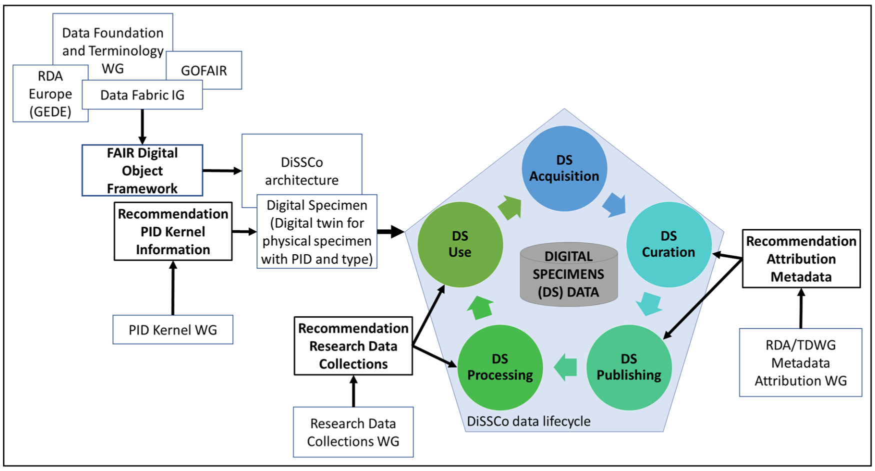
Data Foundation and Terminology WG
RDA Europe (GEDE)
GOFAIR
Data Fabric IG
FAIR Digital Object Framework
DiSSCo architecture
Digital Specimen (Digital twin for physical specimen with PID and type)
Recommendation PID Kernel Information
PID Kernel WG
Recommendation Research Data Collections
Research Data Collections WG
DS Acquisition
DS Use
DS Curation
DIGITAL SPECIMENS (DS) DATA
DS Processing
DS Publishing
DiSSCo data lifecycle
Recommendation Attribution Metadata
RDA/TDWG Metadata Attribution WG

**Figure 2:** Contributions of RDA outputs to the design of data management in the DiSSCo Digital Specimen data infrastructure. The FAIR Digital Object Framework and the Recommendation on PID Kernel Information contribute to the architecture as a whole while the Recommendation on Research Data Collections and Attribution Metadata contribute more explicitly into specific phases of the data lifecycle.

[7] ICEDIG – "Innovation and consolidation for large scale digitisation of natural heritage" – is an EU-funded project (grant agreement number 777483) that aims at supporting the implementation phase of the new Research Infrastructure DiSSCo ("Distributed System of Scientific Collections") by designing and addressing the technical, financial, policy and governance aspects necessary to operate such a large distributed initiative for natural sciences collections across Europe.

[8] We define 'FAIRness' as a characteristic exhibited by an infrastructure (or component thereof) when it achieves and maintains compliance with the FAIR Guiding Principles.

This lifecycle begins with the digitization and acquisition of data from physical specimens – the creation of the Digital Specimens (DS) and Digital Collections that are specific object types with persistent identifiers and attributes. This is the data acquisition phase. These objects then are registered and curated within a repository platform (curation phase). Curated data is published to DiSSCo users and parties external to the infrastructure, as well as directly to other services. DiSSCo will provide services for further processing of data (data processing phase) that can produce new data to be stored within the infrastructure. Finally, the broader research community can use DiSSCo data and can design experiments and analyses acting on the published Digital Specimen and Collection data that produce results (derived data), which in turn can be passed back into DiSSCo for curation, publishing and processing; thus, restarting the lifecycle (DiSSCo DMP, 2019).

## Adoption of Digital Object Architecture

**RDA output from the Data Fabric IG on virtual layer recommendations (RDA DFIG 2018) and the Data Foundation and Terminology WG on the basic vocabulary to apply a standard core data model (RDA DF&T 2015) provide the structure for DiSSCo's data organization model.**

Even though the history behind digital objects goes back to the early days of the Internet (Kahn and Wilensky 2006), the recent rendition has its origin in the RDA's Data Foundation and Terminology (DF&T) WG. From there the discussion has been taken up by the members of the Data Fabric IG (DFIG) together with the C2CAMP[9] initiative, the RDA-Europe Group of European Data Experts (GEDE)[10] and the GOFAIR[11] initiative. These discussions have shaped the current principles of Digital Object Architecture (Sharp 2016) and most recently FAIR Digital Objects (FAIR-DO) (**Figure 3a**) and the FAIR-DO Framework (De Smedt 2020; Wittenburg and Strawn 2019; European Commission 2018). DiSSCo adopts the Digital Object Architecture and the FAIR-DO framework to achieve FAIRness and meet the requirements of the FAIR Guiding Principles for scientific data management (Wilkinson 2016; Lannom et al. 2020; Hardisty et al. 2020).

The main impetus behind adopting FAIR-DOs for Digital Specimens (**Figure 3b**) is to treat the digital representations of physical specimens as atomic items that need individual identification to avoid ambiguity[12] and to collect and anchor core information about the specimens in one place. The Digital Specimens act as the mutable space for the curation of all data derived from and relating to the corresponding physical

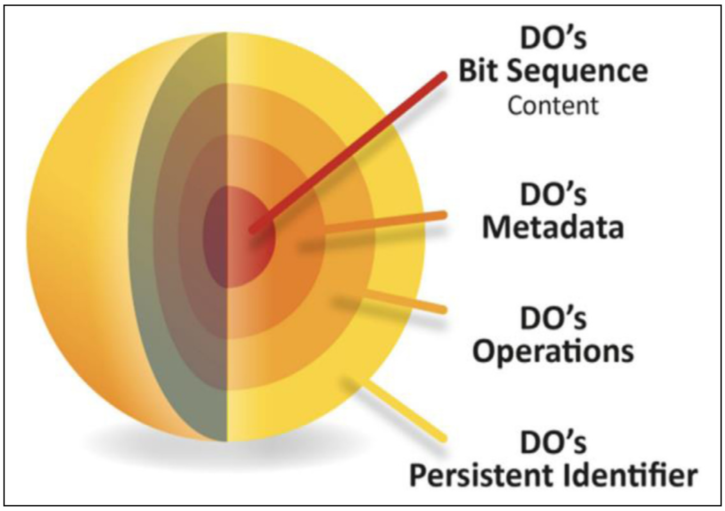

**Figure 3a:** Main components of a digital object – The core of the DO is a bit sequence that is encoding content (data, metadata, software, etc.). This is described by metadata to enable access and for correct interpretation. A persistent identifier uniquely identifies the DO and operations permit the content and metadata to be manipulated. Reproduced with permission (Wittenburg et al. 2019).

[9] C2CAMP: (Cross-continental Collection Access and Management Pilot) https://www.go-fair.org/implementation-networks/overview/c2camp/.

[10] Group of European Data Experts in RDA: https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda.

[11] https://www.go-fair.org/.

[12] The subject of ambiguity looms large when it comes to referring to things by species name (each plant, animal, fossil or rock/mineral specimen typically is labelled with its scientific name if known). Scientific names are immensely useful for taxonomic research and in particular for describing the object. However, they are not usable for disambiguation or by machine-actionable services (see Patterson et al. 2016, Guralnick et al. 2015, Sterner and Franz 2017).

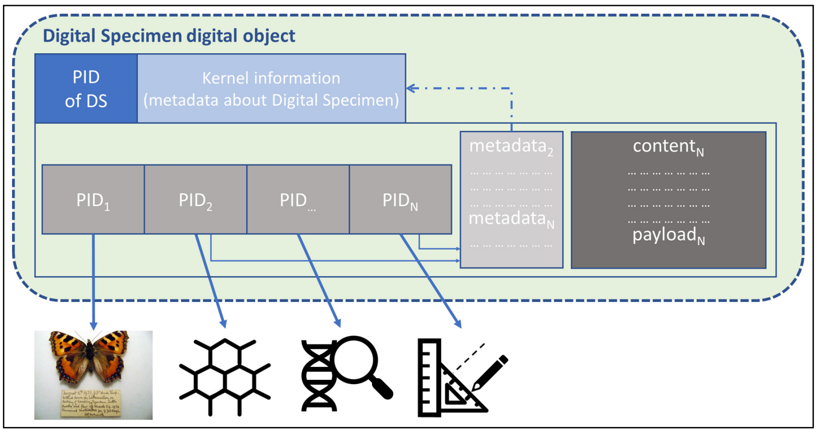

**Figure 3b:** Basic structure of a Digital Specimen (DS). A DS acts as a container for pointers, metadata and embedded content, i.e., information about and derived from the corresponding physical specimen including but not limited to, for example, necessary information about the specimen, image(s), molecular data, genetic sequence data, and morphological measurements.

specimens. Unambiguous persistent identification allows tracking of Digital Specimens in the face of changing location, as well as organization into collections for specific purposes. The data derived from and linked to physical specimens must be easily findable and accessible. They must adhere to open standards with rich machine-comprehensible semantics, as well as conveying context (Stocker 2018) so they are interoperable and widely reusable by both humans and machines. Just being machine-readable (i.e., by linking to ontologies and encoding as RDF or JSON-LD) is insufficient to achieve reusability and, especially for reproducibility of science, provenance, data quality, credit and attribution (Bechhofer et al. 2013).

DiSSCo envisions that persistently and unambiguously identifying these Digital Specimens creates a digital doorway that allows researchers to do more than just find specimens and provide means for the institutions to widen access to the data stored within the NSCs (Webster 2017, Lendemer et al. 2020; Schindel and Cook 2018). DiSSCo expects that adopting the Digital Object Architecture recommended by RDA, and treating Digital Specimens as first-class citizens in that architecture can lead to transformations in working practices of collections-based science and the value chains founded in natural science collections.

## Persistent Identifiers and Kernel Information

**RDA output from the PID Kernel WG on PID Kernel Information (RDA PID KI 2019) provides the capability to elevate a small number of essential attributes of Digital Specimens to the PID record level to enable new machine-actionable services without requiring access to or retrieval of the Digital Specimen objects themselves.**

Identifiers are used in NSCs to identify physical specimens (Güntsch et al. 2017) and organizations such as GBIF are in the forefront of using Digital Object Identifiers (DOI) for datasets, queries and download records (Copas et al. 2019). At the moment though, it is not possible to unambiguously and persistently refer to digital equivalents of a physical specimen. The Digital Specimen and persistent identifier (PID) scheme combination proposed by DiSSCo fills this gap. DiSSCo, with extensive consultation and support from several international stakeholders such as GBIF, Corporation for National Research Initiative (CNRI),[13] International DOI Foundation (IDF),[14] is working towards adopting a Handle-based system (Sun et al. 2003) tuned to the needs of the natural science collections community. Scalability in the tens to hundreds of billions of PIDs (i.e., supporting a huge address space), trust (i.e., accurately maintained by a dedicated and reliable team), persistence over a very long term (i.e.,100-year target) and community governance (i.e., transparent and sustainable business model) are essential requirements to be accommodated. Besides specimens, other things have to be persistently identified. GRID[15] and ROR[16] are used as unique identifiers for institutions and ORCID[17]

[13] https://www.cnri.reston.va.us/.
[14] https://www.doi.org/doi_handbook/7_IDF.html.
[15] https://www.grid.ac/.
[16] https://ror.org/.
[17] https://orcid.org/.

is used for people. These allow to unambiguously link specimens respectively to the collection holding institutions and to the researchers and curators.

Assigning identifiers is the first step towards FAIR data services and ensuring machine actionability of FAIR-DOs. The definition and description of the metadata attributes of the specific digital object and persistent link to all these are the next steps. DiSSCo Data Management Plan recognizes this and thus references the RDA Recommendation on PID Kernel Information (RDA PID KI 2019): "Specific PID Kernel Information profiles and object type definitions must be registered for the Digital Specimen object type and other object types in the well-known Kernel Information profile and Data Type registries" (DiSSCo DMP 2019).

It is clear that a minimal set of information associated with each Digital Specimen should be available to facilitate machine-actionable services and programmatic decisions and delivery of these attributes must work with low latency and in a scalable fashion (Weigel et al. 2020). What is less clear is what these attributes should be or the extent of them, and what makes an optimal kernel information profile. This needs further study.

One use case that can exploit kernel information is submitting large number (millions) of specimen images in long term storage to a workflow for optical character/text recognition (OCR), making the results findable with a full-text search (Cazenave et al. 2019). These images and OCR'd label texts will reside in an ecosystem with millions of other digital objects (also with research artefacts from different domains). Full resolution of each PID might not be feasible in such cases. So for quick machine interpretation processing appropriate kernel information will be vital. A simple kernel information profile example to support this is in **Table 2**.

In this example, "123prefix/uuid-27a9edf63" is the PID of a digital object with several attributes in its particular kernel information profile:

1. Location: URL redirecting to the location of the Digital Specimen object. This URL can resolve to a digital object repository or another landing page and can also provide data serialisation options like JSON-LD.
2. Created: The timestamp when this object was created.
3. Type: A Digital Specimen. Instead of storing the string "Digital Specimen", we refer to a PID in a Data Type Registry which is a resolvable entity with other metadata attached. It tells us the structure of the Digital Specimen object, thus enabling us to parse that.
4. PhysicalSpecimenId: Digital Specimen is a digital twin of a physical specimen, so the identifier of the physical specimen is an important and special attribute for this particular type of digital object. The value here contains the physical specimen identifier as a string.

From a simple machine-actionable point of view, this digital object provides the persistent identifier, points to a type declaration and, provides the physical specimen identifier. Other attributes – currently under discussion in DiSSCo Prepare WP 6 (Technical Architecture and Services provision)[18] – such as scientific names, physical location, version, digitization level/definition, digital object policy, etc.) can also be included in such a profile. These can help to decide whether a Digital Specimen is suitable for the intended operation. For example, an update operation on a Digital Specimen can be adding missing records or fixing incorrect georeference and locality data. This update would be preceded by a search operation that will retrieve incomplete relevant records.

These operations will be part of services envisioned in DiSSCo such as the digitization workflow. At the moment, digitization activities vary from one specimen category to another and between institutions (Cocks et al 2020). We are addressing these challenges within the context of developing openDS (an open specification of Digital Specimen and other related object type definitions essential to mass digitization) and

**Table 2:** Simple example of PID Kernel Information for a Digital Specimen. Example PID: 123prefix/uuid-27a9edf63.

| Attribute | Value Type | Example Value |
|---|---|---|
| Location | url | http://example-dissco-repo/uuid-27a9edf63 |
| Created | date and time | 2019-04-24T11:07:11.771Z |
| Type | type definition | typedef123/DigitalSpecimen |
| PhysicalSpecimenId | string | BMNH:1905.5.30.352 |

[18] https://www.dissco.eu/dissco-prepare-work-programme/.

MIDS (Minimum Information about a Digital Specimen) to establish data standard and common practices.[19] A common understanding of these processes will help us refine how and when the minimum set of metadata that does not change during the lifetime of the object needs to be created and maintained.

Digital Specimens now can become part of a FAIR infrastructure implementation because with kernel information and other metadata, they are findable and accessible. Repository and application services can be built in conjunction with these digital twins as the basis of a Digital Object Architecture. Digital Specimens can be accessed, retrieved and interacted with using standardized communication protocols such as Digital Object Interface Protocol (DOIP[20]) or Hypertext Transfer Protocol (HTTP). Digital Specimens are interoperable because services and systems can determine the attributes that are tied semantically to FAIR vocabularies, and perform operations on them. And lastly, the kernel information profile and other attributes enable accurate and relevant data needed for reproducibility and reusability (for example, publishing the digitized data in a different format or running an experiment using data linked to a specimen).

## Aggregation of digital objects

**RDA output from the Research Data Collection WG on actionable collections and a technical interface specification to enable client-server interaction provide guidelines for how to create meaningful services around the DiSSCo specific digital objects.**

Building on the essential components of Digital Object Architecture and PID scheme, the next step considers how to go beyond the single data objects. "Collection" in the RDA sense means grouping objects together without demanding particular semantics or formats and this grouping should have a unique identifier with well defined actions such as CRUD (Create, Read, Update, Delete) that act on all objects in the group equally.

The NSC community has focused on creating a standard for Collection Descriptions.[21] Furthermore, in NSC and DiSSCo terms, "Collection" has a specific meaning – "A collection is any set of physical things (material/natural objects) or image, audio and video recordings (either analogue or digital) treated together for curative purposes" (Addink et al. 2020). So we need to investigate further to see how commensurate this is with the RDA "Research Data Collection" recommendation (RDA Research Data Collections 2017).

In DiSSCo, a "Digital Collection" is a specific type of digital object acting as a twin for a real-world natural science collection. It is a collection of Digital Specimens, mirroring physical world practice of organizing specimens into specific kinds of collection (zoology, botany, etc.). In the digital world, however, the notion of a collection is far more flexible; insofar as objects can be members of multiple collections simultaneously, even without specific criteria defining membership.

Collections as digital objects will be consumed by services like ELViS (European Loans and Visits System) to facilitate loans and visits transactions and digitization on demand. A Collection Monitoring Dashboard can provide comprehensive overviews of collections across different institutions and disciplines. An extensive set of user stories[22] maps user journeys for activities such as searching for collections and specimens, requesting loans, reviewing loan requests, generating reports on loans and visits and collection usage. For each of these steps and services, the role of collection as a digital object is crucial (**Figure 4**).

Following the RDA recommendation (RDA Research Data Collections 2017), the digital collection as an entity with a persistent identifier (e.g., the digital object for the Mammal Collection at a museum) will support different operations (such as retrieve ordered or filtered list) and specific properties (essential information such as which museum, how it is stored), and membership information (e.g., specimens that are in the mammal collection). Beside ELViS, other e-services are envisaged to be implemented on collections of specific data types; for example, involving automated machine learning and computer vision (Livermore and Cubey 2019).

One of the challenges the DiSSCo architecture design work needs to address is the difficulty of understanding how data as a bundled package moves and is used across different practical situations in science, industry, policy making and public discourse. Data can be decontextualized and then recontextualized in novel situations to become meaningful beyond their original context of production (Leonelli 2016). Digital

[19] Work on the first of these specifications, openDS is at an early stage while work on the MIDS standard is more mature, having recently (September 2020) been taken up by TDWG – the community group responsible for Biodiversity Information Standards in a new Task Group on Minimum Information about a Digital Specimen (MIDS), https://www.tdwg.org/community/cd/mids/.

[20] https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf.

[21] Collection Description interests group in Biodiversity Information Standard (TDWG): https://github.com/tdwg/cd.

[22] DiSSCo user stories in github: https://github.com/DiSSCo/user-stories/.

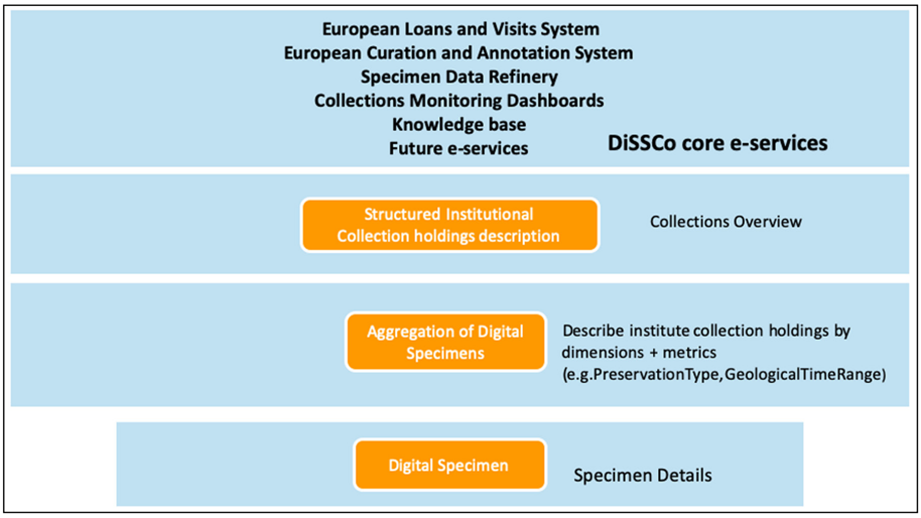
European Loans and Visits System
European Curation and Annotation System
Specimen Data Refinery
Collections Monitoring Dashboards
Knowledge base
Future e-services
DiSSCo core e-services
Structured Institutional Collection holdings description
Collections Overview
Aggregation of Digital Specimens
Describe institute collection holdings by dimensions + metrics (e.g.PreservationType,GeologicalTimeRange)
Digital Specimen
Specimen Details

**Figure 4:** Building blocks of DiSSCo e-services start with individual objects (represented digitally through Digital Specimens), collections and collections overview.

Specimens as atomic entities and the flexibility to organize them and other object types into different kinds of machine-actionable digital object container (i.e., 'collection' in the RDA sense) will help in facilitating yet to be imagined uses.

## Metadata attribution and use of PROV entities

**Output from the joint RDA/TWDG metadata attribution WG on standardized metadata for attributing work and tracking provenance in the curation and maintenance of research collections guides how attribution details can be preserved in DiSSCo digital objects.**

DiSSCo's Data Management Plan (DiSSCo DMP 2019) highlights the importance of the provenance of specimens, their digitization and change history, annotation, curation, and usage. These histories must be maintained consistently through the lifetime of the DiSSCo infrastructure. Recording activities of human and machine agents during data curation and processing phases is essential to FAIR implementation. Groom et al. (2019:12) highlight the importance of attributing people behind the specimens in a standard fashion: "*Many people can be associated with a specimen: the collector, curator, determiner, annotator, mounter, transcriber, digitizer, imager and georeferencer. For many reasons, these people are important to science. Knowing the person gives a degree of credibility to the specimen and its identity. The biographical data of the people can not only help validate data, but also credit the people for the work they have done*".

Three important information elements capture this detail and can be used as the means for attributing work: the agent performing the activity, the activity (or action) they perform, and the digital or physical object (entity) they are curating/processing. The joint RDA/TDWG Attribution WG Metadata recommendation (RDA/TDWG Attribution Metadata 2018), which uses W3C's PROV data model, makes it easy to implement this in the FAIR-DO context. In our design, we are planning to store provenance as another digital object type linked to the object on which work is performed. This will create a common curation space linked to different types of digital objects, other data classes and services. One current proposal to materialize this is through the European Curation and Annotation System (ECAS) e-service to serve as a community curation service.

One of the example use cases in the RDA recommendation is relevant:

> "*Sergey (a museum curator) recurates a jar containing multiple specimens. Each specimen is removed from the jar and individually mounted. Sergey then examines the specimen and jar label, and enters a new record into the collections management database. He uses the data in the new record to generate a new label to attach to the physical specimen.*
> *Sergey also, in the process of recurating one of the specimens, discovers a new species.*
> *He describes the new species, and uses the species description to publish a journal paper.*
> *Sergey should receive attribution for:*
>
> - *recurating the physical specimens*

- *describing the new species*
- *authoring the journal article*
- *entering the specimen into the collections management database*
- *generating a label for the re-curated specimen."*

As is evident, even within a single workflow, data can travel from a collection management database to a journal where different systems, standards, and application programming interfaces (API) are involved. These five attributions need to be captured in a standard way to be part of the Digital Specimen data when different operations are performed in multiple contexts.

Leonelli (2016: 188–189) using the example of model organism biology, makes the point that to support the scientists, we must understand the processes behind successful empirical research. Often, policymakers and funders predominantly understand research as products instead of processes. Metadata attribution and use of PROV entities provide technical foundation to bring these processes to the forefront of supporting and sustaining a research infrastructure.

## FAIR Data Maturity Model

**Output from the RDA FAIR data maturity model WG (RDA FAIR Data Maturity Model 2020) provides guidelines and specifications to assess the DiSSCo FAIR implementation plan.**

DiSSCo's Data Management Plan (DiSSCo DMP 2019, Appendix E) provides a summary statement of DiSSCo's implementation of the FAIR guiding principles. The indicators, priority levels and evaluation methods described by the FAIR Data Maturity Model (DMM) WG (RDA FAIR Data Maturity Model 2020) were not available during the preparation of the DiSSCo DMP. However, the output is an essential tool for future periodic evaluation of the DMP and FAIR implementation.

As DiSSCo data infrastructure is FAIR by design, the essential indicators in the DMM are thus addressed. At the time of writing this article, DiSSCo is in maturity level "2" ("under consideration or in planning phase") for all the essential indicators. The DMM also decomposes texts of the FAIR principles to provide further granularity. For instance, the RDA output provides two indicators for FAIR principle F1[23] (one for persistent identifier and one for globally unique identifier). DiSSCo DMP addresses F1 as such: *"A handle is issued to each object published in or by DiSSCo, allowing the object data to be found regardless of its location"*. Due to our design choice of FAIR Digital Object, DiSSCo addresses both the persistency and uniqueness aspect of F1. For FAIR principle R1, the DMM indicator is: "*Plurality of accurate and relevant attributes are provided to allow reuse"* which is based on "R1: *(Meta)data are richly described with a plurality of accurate and relevant attributes*". For R1, DiSSCo DMP states:

- Each object contains a minimum of mandatory terms consistent with its formal object type definition, with the possibility to include optional additional terms and enrichments as necessary.
- In the case of Digital Specimen and Digital Collection object types, the minimum of mandatory terms corresponds to the object's classification as representing a specific level of digitization according to (respectively) the Minimum Information standard for Digital Specimens (MIDS) and the Minimum Information standard for Digital Collections (MICS).

Implementation of MIDS in the digitization process will ensure that enough data are captured, curated and published to make it reusable and thus creating "plurality of accurate and relevant attributes". As we progress along from design phase to pilot and then implementation, the DMM indicators and evaluation methods can help DiSSCo to create tailored assessment but at the same time focus on FAIR convergence for cross-disciplinary interoperability (Sustokova et al. 2020). Similarly, the other indicators mentioned in the DMP is commensurable with the RDA DMM framework.

## Conclusions

In this paper, we have presented how the RDA outputs can be used to create building blocks for research infrastructure architectural design decisions towards FAIR compliance. For DiSSCo, e-services such as ELViS, designed around the concept of Digital Specimens are planning to improve access to natural science collections across Europe. Aggregation of these Digital Specimens through Digital Collections will enable moni-

[23] F1: (Meta)data are assigned a globally unique and persistent identifier.

toring tools like CMD to provide collections overview and reports that are immensely beneficial to track and assess scientific usage of the collection. The RDA outputs are not just for the access/use part of the data lifecycle. Data enhancement, annotation (using the planned Specimen Data Refinery) and community curation (using the European Curation and Annotation System) are building blocks for the research infrastructure vision of DiSSCo that all depend on these recommendations. Along with the different building blocks, the outputs also highlight the importance of data standards and common practices which have already been discussed in the ICEDIG project (2018–2020) and are currently being further studied in DiSSco Prepare (2020–2023).

The ideas expressed here are still in the design and/or conception stage and need to be further fleshed out to support the DiSSCo implementation and construction phase. Some of the RDA outputs are also similar in their conceptual nature and thus organizing workshops, and technical hackathons through RDA can help DiSSCo to clarify further, test and refine the concepts. DiSSCo experts regularly participate in RDA and collaboration with other disciplines through RDA can also provide learning opportunities and help us identify potential issues and risks in our concepts. There are other outputs – such as the outputs of the PID Information Types WG (RDA PID Information Types 2015) and the Data Type Registries WG (RDA DTR 2015) – that we are still exploring.

The RDA recommendations and the broader global expertise represented therein enable us to design and build a robust, FAIR Digital Object based data infrastructure. We envision that this new infrastructure will be essential in supporting the next phase in the digital transformation of collections-based science to widen access and better enable the production of data and knowledge about the 1.5 billion physical specimens in the European natural science collections.

## Acknowledgements

This paper was supported by the RDA Europe 4.0 project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777388. The DiSSCo Prepare project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871043.

## Competing Interests

The authors have no competing interests to declare.

## References

**Addink, W,** et al. 2020. Advancing the Catalogue of the World's Natural History Collections – 10 recommendations from DiSSCo. DOI: https://doi.org/10.5281/zenodo.3949839

**Agosti, D,** et al. 2019. Biodiversity Literature Repository (BLR), a repository for FAIR data and publications. *Biodiversity Information Science and Standards*. DOI: https://doi.org/10.3897/biss.3.37197

**Bechhofer, S,** et al. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2): 599–611. DOI: https://doi.org/10.1016/j.future.2011.08.004

**Besnard, G,** et al. 2016. Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biological Journal of the Linnean Society*, 117(1): 71–82. DOI: https://doi.org/10.1111/bij.12494

**Blagoderov, V, Kitching, IJ, Livermore, L, Simonsen, TJ** and **Smith, VS.** 2012. No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys*, 209: 133. DOI: https://doi.org/10.3897/zookeys.209.3178

**Cazenave, N, Béchard, L** and **Rouchon O.** 2019. Digitisation infrastructure design for EUDAT/CINES. *ICEDIG Deliverable 6.2*. DOI: https://doi.org/10.5281/zenodo.3364533

**Chapman, A,** et al. 2020. Developing standards for improved data quality and for selecting fit for use biodiversity data. *Biodiversity Information Science and Standards*, 4: e50889. DOI: https://doi.org/10.3897/biss.4.50889

**Cocks, N, Livermore, L, Smith, V** and **Woodburn, M.** 2020. Technical capacities of digitisation centres within ICEDIG participating institutions. *Research Ideas and Outcomes*, 6: e55522. DOI: https://doi.org/10.3897/rio.6.e55522

**Cook, JA,** et al. 2020. Integrating Biodiversity Infrastructure into Pathogen Discovery and Mitigation of Emerging Infectious Diseases. *BioScience*, 70(7): 531–534. DOI: https://doi.org/10.1093/biosci/biaa064

**Copas, K, Noesgaard, D** and **Schigel, D.** 2019. Crediting the reuse and impact of free, FAIR and open biodiversity data through DOI citations and event tracking. *AGUFM*, 2019: IN21A-06.

**De Smedt, K, Koureas, D** and **Wittenburg, P,** 2020. FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. *Publications*, 8(2): 21. DOI: https://doi.org/10.3390/publications8020021

**Díaz, S,** et al. 2016. The global spectrum of plant form and function. *Nature*, 529(7585): 167–171. DOI: https://doi.org/10.1038/nature16489

**DiSSCo DMP.** 2019. Provisional Data Management Plan for the DiSSCo infrastructure. DOI: https://doi.org/10.5281/zenodo.3532937

**European Commission.** 2018. Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR data. *Luxembourg Publication Office of the European Union, Luxembourg*, 78. DOI: https://doi.org/10.2777/1524

**Grobe, P,** et al. 2019. From Data to Knowledge: A semantic knowledge graph application for curating specimen data. *Biodiversity Information Science and Standards*. DOI: https://doi.org/10.3897/biss.3.37412

**Groom, Q, Dillen, M, Hardy, H, Phillips, S, Willemse, L** and **Wu, Z,** 2019. Improved standardization of transcribed digital specimen data. *Database*, 2019. DOI: https://doi.org/10.1093/database/baz129

**Güntsch, A,** et al. 2017. Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *Database*, 2017. DOI: https://doi.org/10.1093/database/bax003

**Guralnick, RP,** et al. 2015. Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys*, 494: 133. DOI: https://doi.org/10.3897/zookeys.494.9352

**Hardisty, A.,** et al. 2020. Conceptual design blueprint for the DiSSCo digitization infrastructure-DELIVERABLE D8. 1. *Research Ideas and Outcomes*, 6. DOI: https://doi.org/10.3897/rio.6.e54280

**Hardisty, A** and **Roberts, D.** 2013. A decadal view of biodiversity informatics: challenges and priorities. *BMC ecology*, 13(1): 16. DOI: https://doi.org/10.1186/1472-6785-13-16

**Hedrick, BP,** et al. 2020. Digitization and the future of natural history collections. *BioScience*, 70(3): 243–251. DOI: https://doi.org/10.1093/biosci/biz163

**Hobern, D,** et al. 2019. Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity data journal*, 7. DOI: https://doi.org/10.3897/BDJ.7.e33679.suppl10

**Kahn, R** and **Wilensky, R,** 2006. A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2): 115–123. DOI: https://doi.org/10.1007/s00799-005-0128-x

**Kays, R,** et al. 2020. An empirical evaluation of camera trap study design: How many, how long and when?. *Methods in Ecology and Evolution*. DOI: https://doi.org/10.1111/2041-210X.13370

**Kays, R, McShea, WJ** and **Wikelski, M.** 2020. Born-digital biodiversity data: Millions and billions. *Diversity and Distributions*, 26(5): 644–648. DOI: https://doi.org/10.1111/ddi.12993

**Lannom, L, Koureas, D** and **Hardisty, AR.** 2020. FAIR data and services in biodiversity science and geoscience. *Data Intelligence*, 2(1–2): 122–130. DOI: https://doi.org/10.1162/dint_a_00034

**Lendemer, J,** et al. 2020. The extended specimen network: A strategy to enhance US biodiversity collections, promote research and education. *BioScience*, 70(1): 23–30. DOI: https://doi.org/10.1093/biosci/biz165

**Leonelli, S,** 2016. *Data-centric biology: A philosophical study*. University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226416502.001.0001

**Lewis, KP, Vander, Wal, E** and **Fifield, DA.** 2018. Wildlife biology, big data, and reproducible research. *Wildlife Society Bulletin*, 42(1): 172–179. DOI: https://doi.org/10.1002/wsb.847

**Lister, AM** and **Climate Change Research Group.** 2011. Natural history collections as sources of long-term datasets. *Trends in ecology & evolution*, 26(4): 153–154. DOI: https://doi.org/10.1016/j.tree.2010.12.009

**Livermore, L** and **Cubey, R.** 2019. Specimen Data Refinery: A landscape analysis on machine learning, computer vision and automated approaches to capture specimen metadata. *Biodiversity Information Science and Standards*. DOI: https://doi.org/10.3897/biss.3.37647

**Martin, P, Chen, Y, Hardisty, A, Jeffery, K** and **Zhao, Z.** 2017. Computational Challenges in Global Environmental Research Infrastructures. In: *Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities*, Chabbi, A and Loescher, HW (eds.). CRC Press ISBN 9781498751315. DOI: https://doi.org/10.1201/9781315368252

**Mons, B,** et al. 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37(1): 49–56. DOI: https://doi.org/10.3233/ISU-170824

**Nachman, MW.** 2013. Genomics and museum specimens. *Molecular Ecology*, 22(24): 5966–5968. DOI: https://doi.org/10.1111/mec.12563

**Nelson, G** and **Ellis, S.** 2019. The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B*, 374(1763): 20170391. DOI: https://doi.org/10.1098/rstb.2017.0391

**Nieva de la Hidalga, A, Hardisty, A, Martin, P, Magagna, B** and **Zhao, Z.** 2020. The ENVRI Reference Model. In *Towards Interoperable research infrastructures for environmental and earth sciences – A reference model guided approach for common challenges*, Zhao, Z and Hellstrom, M (eds.). LNCS 12003, 61–81. DOI: https://doi.org/10.1007/978-3-030-52829-4_4

**Page, R.** 2016. Towards a biodiversity knowledge graph. *Research Ideas and Outcomes*, 2. DOI: https://doi.org/10.3897/rio.2.e8767

**Patterson, D, Mozzherin, D, Shorthouse, DP** and **Thessen, A.** 2016. Challenges with using names to link digital biodiversity information. *Biodiversity Data Journal*, 4. DOI: https://doi.org/10.3897/BDJ.4.e8080

**RDA DTR.** 2015. Data Type Registries working group output. DOI: https://10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

**RDA DF&T.** 2015. Data Foundation and Terminology Work Group Products. *Research Data Alliance.* DOI: http://doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF

**RDA DFIG.** 2018. Data Fabric Interest Group; Summary of Virtual Layer Recommendations. *Research Data Alliance.* DOI: http://doi.org/10.15497/RDA00026

**RDA FAIR Data Maturity Model.** 2020. FAIR Data Maturity Model: specification and guidelines. *Research Data Alliance.* DOI: http://doi.org/10.15497/RDA00050

**RDA PID KI.** 2019. RDA Recommendation on PID Kernel Information. *Research Data Alliance.* DOI: https://doi.org/10.15497/RDA00031

**RDA PID Information Types.** 2015. Final deliverable. *Research Data Alliance.* DOI: https://doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786

**RDA Research Data Collections.** 2017. Recommendation on Research Data Collections. *Research Data Alliance.* DOI: https://doi.org/10.15497/RDA00022

**RDA/TDWG Attribution Metadata.** 2018. Final Recommendations. *Research Data Alliance.* DOI: http://doi.org/10.15497/RDA00029

**Schindel, DE** and **Cook, JA.** 2018. The next generation of natural history collections. *PLoS Biology*, 16(7): e2006125. DOI: https://doi.org/10.1371/journal.pbio.2006125

**Senderov, V,** et al. 2018. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. *Journal of biomedical semantics*, 9(1): 5. DOI: https://doi.org/10.1186/s13326-017-0174-5

**Sharp, C.** 2016. Overview of the digital object architecture (DOA). *An Internet Society Information Paper, Internet Society.* Retrieved from https://www.internetsociety.org/resources/doc/2016/overview-of-the-digital-object-architecture-doa/.

**Smith, V,** et al. 2019. SYNTHESYS+ Abridged Grant Proposal. *Research Ideas and Outcomes*, 5: e46404. DOI: https://doi.org/10.3897/rio.5.e46404

**Sterner, B** and **Franz, NM.** 2017. Taxonomy for humans or computers? Cognitive pragmatics for big data. *Biological Theory*, 12(2): 99–111. DOI: https://doi.org/10.1007/s13752-017-0259-5

**Stocker, M,** et al. 2018. Curating Scientific Information in Knowledge Infrastructures. *Data Science Journal*, 17(21): 1–16. DOI: https://doi.org/10.5334/dsj-2018-021

**Sun, S, Lannom, L** and **Boesch B.** 2003. Handle System Overview, RFC 3650. DOI: https://doi.org/10.17487/rfc3650

**Sustkova, HP, Hettne, KM, Wittenburg, P, Jacobsen, A, Kuhn, T, Pergl, R, Slifka, J, McQuilton, P, Magagna, B, Sansone, SA** and **Stocker, M.** 2020. FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. *Data Intelligence*, 2(1–2): 158–170. DOI: https://doi.org/10.1162/dint_a_00038

**Webster, MS.** (ed.) 2017. *The extended specimen: emerging frontiers in collections-based ornithological research.* CRC Press.

**Weigel, T,** et al. 2020. Making data and workflows findable for machines. *Data Intelligence*, 2(1–2): 40–46. DOI: https://doi.org/10.1162/dint_a_00026

**Wilkinson, M, Dumontier, M, Aalbersberg, I,** et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18

**Wittenburg, P,** et al. 2019. Digital objects as drivers towards convergence in data infrastructures. *Technical paper.* DOI: http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11

**Wittenburg, P** and **Strawn, G.** 2019. Commenting on "Digital Object" Aspects. DOI: http://doi.org/10.23728/b2share.2317b12321764f669c92ebbcf7518164

**How to cite this article:** Islam, S, Hardisty, A, Addink, W, Weiland, C and Glöckler, F. 2020. Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. *Data Science Journal,* 19: 50, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2020-050

**Submitted:** 17 July 2020 **Accepted:** 17 November 2020 **Published:** 14 December 2020

**Copyright:** © 2020 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

]u[ *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS