ARPHA Preprints

PENSOFT

Project Report

Author-formatted document posted on 30/05/2023

Published in a RIO article collection by decision of the collection editors.

DOI: https://doi.org/10.3897/arphapreprints.e107168

Deliverable D7.1 Architecture Design for a pan-European PID system for Digital Specimens

Wouter Addink, Sharif Islam, Mathias Dillen, Anton Güntsch, Soulaine Theocharides

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

Architecture Design for a pan-European PID system for Digital Specimens

Deliverable D7.1

31 October 2022

Authors

Wouter Addink[1], Sharif Islam[1], Mathias Dillen[2], Anton Güntsch[3], Soulaine Theocharides[1]

1: Naturalis Biodiversity Center

2: Meise Botanic Garden

3:Freie Universität Berlin

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Architecture Design for a pan-European PID system for Digital Specimens
Deliverable n°:	D7.1
Nature of the deliverable:	Report
Dissemination level:	Public
WP responsible:	WP7
Lead beneficiary:	Naturalis
Citation:	Addink, W., Islam, S., Dillen, M., Güntsch, A. & Theocharides S. (2022). *Architecture Design for a pan-European PID system for Digital Specimens*. Deliverable D7.1 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month n° 18
Actual submission date:	31 October 2022

Deliverable status:

Version	Status	Date	Author(s)
1.0	Final	Oct 2022	W Addink, Naturalis

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

Table of contents

Preface

This system design document outlines the planned architecture with various components of a Persistent Identifier (PID) system for Digital Specimens. The rationale for such a system and the decision to use DOI and be part of the International DOI Foundation (DOIF) are elaborated elsewhere. This document first provides a primer to the PID technologies, highlights a few terms for better readability, then outlines the components and lists requirements.

For more background information, the reader should consult the provisional DiSSCo Data Management Plan and DiSSCo user stories. The 7.1 task team also consulted the PID architecture for the EOSC and the PID policy for the European Open Science Cloud (EOSC). Even though BiCIKL is a European project, an informal task group has been created with global infrastructure stakeholders (GBIF, iDigBio, ALA, BCoN, iDigBio, Pensoft Publishers Ltd, LifeWatch ERIC, Biodiversity Heritage Library) to further investigate the possibility of creating a Registration Agency (RA) for Digital Specimen PIDs. PID infrastructure providers such as DataCite and CNRI have also been consulted. Recommendations, agreements, and conclusions coming out of the above global task group may influence the final implementation of the pan-European PID system.

Summary

Persistent Identifier (PID) systems are the foundation for achieving the FAIR Guiding Principles ("findable, accessible, interoperable and reusable"). As FAIR data and connecting different data classes (i.e. specimens, genomics, observations, taxonomy and publications) are essential aspects of the BiCIKL project, we need a PID system at least at the European level to create and maintain identifiers for the digital representation of specimens and samples, called Digital Specimens (DS) (Hardisty et al. 2022). The PID system provides the mechanism to ensure that identifiers are globally unique, persistent and resolvable. This system should also manage associated metadata, facilitate provenance, enable discovery, manage states and the life cycle of the PID, link to other derived data and digital content, and allow content providers to enforce metadata constraints. For the successful provision of a PID system, this design document has been created to guide us during the implementation and operation phases. The document is based on an earlier milestone (MS28) that was used for discussion and evaluation with potential end-users.

List of terms and abbreviations

Term	Description
Registration Agency (RA)	An entity that provides services to Registrants (see below). These services include — allocating DOI name prefixes, registering DOI names and providing the necessary infrastructure to allow Registrants to declare and maintain metadata and state data. In order to become an RA, the

	organisation must be a member of the International DOI Foundation.
Registrant	A client and customer of an RA. This can be any organisation that wishes to uniquely identify entities using the DOI system. Registrants can register DOI names with a Registration Agency. For example, a museum part of the DiSSCo network can be a Registrant.
PID User/Consumer	An agent (human or machine) that interacts with the PID service. Example of PID users/consumer: Collection-owning institutions, Collection managers and Curators, Digitization systems, Researchers and research workflows, Linked data infrastructure, Journals and publishers, Individuals and community/citizen scientists.
Local Handle Service (LHS)	A Local Handle Service (LHS) can be thought of as a service that brokers and stores handle data. “Local” here refers to the namespace part of the naming authorities, not the network topology. Each naming authority defines a “local” namespace that includes all the handles under that authority. A “prefix” is assigned to a naming authority. For example “20.” (20 dot) is the naming authority prefix for CNRI, ”21.” are for handles issued by the German Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), “10.” are for the International DOI Foundation.
Digital Object and FAIR Digital Object	A Digital Object (DO) is an abstraction of data using identification and minimal elements which are technology and implementation agnostic (Kahn and Wilensky 2006). As FAIR and importance of PID for data became prominent around 2016-2017, it was evident that DOIs can be used as mechanism for FAIR implementation (Schwardmann 2020) thus the term FAIR Digital Object (FDO) combines FAIR and Digital Object to provide a conceptual and implementation framework for FAIR data system and service. For more information go to FAIR Digital Objects Forum.
EOSC	The ambition of the European Open Science Cloud (EOSC) is to develop a “Web of FAIR Data and services” for science in Europe. EOSC will be a multi-disciplinary environment where researchers can publish, find and re-use data, tools and services, enabling them to better conduct their work.

1. The Handle System and Digital Object Identifiers

Handle System Architecture

The Handle System is an infrastructure for assigning, managing, and resolving persistent identifiers. It adheres to a distributed architecture, enabling independent servers to store identifiers of digital resources. These identifiers can be resolved globally into the information necessary to locate and access the resources. Records associated with a given handle can be changed to reflect the current state of the resource without changing the identifier itself. This allows the name of an item to persist over changes to its location or other state information.

A Handle is a string that consists of two components: a prefix and a suffix. The prefix is assigned by a global authority and is usually associated with a specific organisation. The suffix must be unique under that prefix. Typically, an organisation establishes one Local Handle Service for each prefix it is assigned. In theory, there is no defined limit on the length of either prefix or suffix and all Unicode characters are supported. Handles are not case-sensitive.

The Handle System has two levels of hierarchy: the Global Handle Registry (GHR) and the Local Handle Services (LHS). The GHR contains records of each registered prefix. When a Handle is resolved, the GHR uses the prefix to determine the correct LHS, and redirects the request accordingly.

Digital Object Identifiers System

The Digital Object Identifier (DOI) system provides actionable, persistent, and interoperable identifiers for entities, such as physical objects (e.g. specimens, books), digital objects (e.g. digital specimens, PDFs), and abstract objects (e.g. taxa, ontological concepts). Through a combination of technical infrastructure and regulations, the DOI system has assigned identifiers for approximately 275 million objects to-date[1]. The number of identifiers we will need for digital specimens is much higher – GBIF currently holds about 200M biological specimen records and there are an estimated 1.5B specimens in Europe and over 3B worldwide. Also, in DiSSCo, there is a plan to have DOIs for the specimen media files, which may be multiple per specimen. This has been discussed with the DOI Foundation members and CNRI. They see no issues with the volume and performance or shifting the focus of the things represented with DOIs. DataCite also does not envision issues during these DOI registration processes. A pilot activity carried out by DiSSCo demonstrated that large numbers of Handles can be minted and updated in bulk, although not with the default Handle server software setup provided by CNRI. A future pilot with DataCite will need to show the feasibility of the DataCite infrastructure for such a large number of digital specimens. DataCite is confident about this but GBIF experienced some performance issues with DOIs minted through DataCite in the past.

The DOI system relies on the Handle System's infrastructure to resolve the PIDs so all DOIs are essentially Handles and are thus resolvable through the Handle System. However, not all

[1] https://www.doi.org/factsheets/DOIKeyFacts.html

Handles are DOIs. The DOI system provides additional value along with the identifier registration. The DOI framework adds metadata requirements to ensure interoperability between DOIs, and the DOI social infrastructure and regulations ensure persistence. It also provides a framework to permit multiple resolutions, allowing one persistent identifier to resolve to multiple locations depending on the context.

Because of the additional services provided by the DOI system, DOIs will be the PIDs of choice for Digital Specimens and associated objects (such as images or audio files). We will have other types of Digital Objects that will not require DOI level support and social guarantee of persistence[2] (based on the membership through the DOI Foundation). For example, annotations, workflows, vocabularies, instruments etc. These Digital Objects might or might not need to be globally resolvable but will certainly need to be linked within various stages of data processing and analysis. For those objects, we will use a Handle instead. The persistence of these handle-based PIDs for instance can be ensured through DiSSCo's policies.

Registration Agencies

Registration Agencies (RAs) are member organisations of the International DOI Foundation (IDF) and can be considered "modules" of the DOI system. RAs offer services for registration of prefixes and individual DOI names to clients (Registrants) who wish to register DOIs.
In order to register DOI names, DiSSCo must establish a relationship with a RA. This may be done through one of the following options:

1. Create a new RA formed by an international coalition of stakeholders providing digital specimen infrastructure or related data objects. The stakeholders are both partners and Registrants to the RA.
2. DiSSCo becomes an RA. Infrastructure is hosted by DiSSCo and used internationally.
3. Integrate DiSSCo infrastructure with DataCite, an existing RA. Providers of digital specimens or related objects will become clients. A draft proposal for a pilot partnership with DataCite is under discussion with the international stakeholders task group.

Option 1 is considered good practice, as the community is directly represented in the RA. Also it gives full control over the DOI names and metadata schemas, is the most cost-effective (no third-party involved), and provides the best way of support for the community. However, since digital specimens are community-owned, and created by the DiSSCo infrastructure, the client would be DiSSCo, which would not need much support. This may be different in other parts of the world where collection holding institutions may want to create digital specimens themselves in the absence of a digital specimen providing infrastructure. This option requires resources from the stakeholders to maintain PID infrastructure and related services, as well as consensus among stakeholders regarding design decisions for the PID infrastructure.

Option 2 is possible, but it is not in DiSSCo's scope to provide an international RA. DiSSCo could only provide DOIs for digital specimens from Europe, but an international service would

[2] https://www.doi.org/factsheets/DOIHandle.html

be preferable. This option should only be considered if the alternative options fail.

Option 3 is only possible if digital specimens and related objects can have their own metadata schemas under control of the digital specimen stakeholders; the current DataCite metadata schema would not be suitable. Discussions with Matt Buys indicated that DataCite seems willing to support this. This option could benefit from the already established relationship between GBIF and DataCite but may also make governance more complex. It could also benefit from direct integration with existing DataCite infrastructure: DataCite may maintain the PID infrastructure and related services. Since a third party is involved (DataCite), it is probably more expensive than Option 1, and the setup of DataCite where clients mint the DOI names may have some effects on the format of the DOI strings and its metadata.

1.1. DOI Names

When combined with its prefix, the DOI suffix renders a DOI name that is globally unique. It is the responsibility of the RA to ensure DOI suffixes are locally unique under the RA's prefix. There are a few restrictions in place for DOI suffixes in terms of format or permitted characters. The following section outlines the restrictions we have put in place to ensure DOI names that meet our needs.

Allowed Characters: For web friendliness, we should restrict the allowed characters in a Digital Specimen DOI (not all UTF8). [A-Z;0-9] seems to be a good choice. We have elected to use Base 32 for our suffixes. Base 32 is a textual 32-symbol notation for expressing numbers in a form that can be conveniently and accurately transmitted between humans and computer systems. It seems a good choice for our needs and is also suggested in a blog post by Martin Fenner. The advantages of creating identifiers this way are that it makes the identifier:

- Human readable and machine readable.
- Compact. Humans have difficulty in manipulating long strings of arbitrary symbols.
- Error resistant. Entering the symbols must not require keyboarding gymnastics.
- Pronounceable. Humans should be able to accurately transmit the symbols to other humans in verbal conversations.

There are implementations for DOI generation available already: Cirneco: command-line client for DataCite MDS and https://github.com/eawag-rdm/doigenerator

Checksum: Often a checksum is added (one or two check symbols) for identifiers with opaque strings. The checksum method can help detect wrong-symbol and transposed-symbol errors and allow for the early detection of transmission and entry errors. For instance, ROR uses leading 0 followed by six characters (excludes I, L, O) and a 2-digit checksum based on the Crockford base-32 URL library and ISO-7064. EIDR uses ISO 7064 with a MOD 37,36 check character. ORCID uses an ISO/IEC 7064:2003, MOD 11-2 checksum. We asked DOI Foundation members and DiSSCo Technical Advisory Board members for advice and their opinion was that the disadvantages of including a checksum (creating a longer string that is more difficult to remember for humans) seem to outweigh the benefits. The chance of transmission and entry errors is very low but adding a checksum may help humans with avoiding typing a wrong PID.

Length: The length of the suffix is dependent on the number of identifiers needed. For

processing, it is preferable to have fixed length strings. For 3 billion digital specimens, we will need 7 symbols (32^7 = 3.43597384×10^10 options). For all other objects 8 symbols would be sufficient (1.09951163×10^12 options). DOI Foundation members strongly advised us to randomly generate the DOIs (not in sequential order) to prevent any assumptions on what will be the next DOI to be minted. To leave room for an optional checksum character we plan to use 9 symbols. We will use these for both the digital objects that get a handle and that get a doi, so for two prefixes. This allows for easy migration to DOIs for more of the digital objects in the future if there is a need for that.

Format: For human readability, we divide the string into manageable parts with a separator when it is more than six characters. EIDR uses a hyphen, e.g. 10.5240/52D6-86C5-C5EB-A682-ECC0-H. The dash and minus signs can create confusion; in that case, a dot may be a better choice. However, if we are going for Base 32 then we should use Hyphens (-), since hyphens are ignored during Base32 decoding. In conclusion, we will use the following format for the Digital Specimen DOI names: https://doi.org/10.22/AB1-2LM-32R

2. Current approaches/practises

There are different types of identifiers currently in use in the biodiversity informatics and natural science collection community. CETAF recommends using stable identifiers (URI based approach) to “redirect users and systems to images, websites and metadata of the physical objects and to integrate them with the semantic web” (see also Güntsch et. al 2017). Other non URI based identifiers such as Darwin Core Triplet (concatenation of values for institution code, collection code, and catalogue number) provide a simple mechanism for institutions to label physical specimens but introduce other issues (Guralnick et. al 2014). The Global Biodiversity Information Facility (GBIF) currently provides DOIs for species occurrences datasets. To identify and track individual specimens, GBIF currently makes use of the unique occurrence IDs that are provided via the data publishers (these records often include CETAF stable identifiers or Darwin Core triplets). GBIF also provides INSDC sequence records (the dataset has a DOI) not associated with environmental sample identifiers or host organisms. This kind of GBIF data links to the accession identifiers of the sequence databases (example: https://www.ebi.ac.uk/ena/browser/api/embl/LC640119). Along with specimen level identifiers, institution and collection level identifiers such as ROR (Research Organization Registry) create other challenges for linking. We also need to consider discipline specific practices and idiosyncrasies (see Sabaj 2020 for a detailed study on ichthyology and herpetology collections). Several journals provide recommendations for material citation (see Material Citations Formatting Guide from the *European Journal of Taxonomy)*. Zenodo provides DOIs for uploaded datasets and also supports DOIs for uploads of a “physical object” type which can be used for specimens. All these approaches are a step towards FAIR implementation and provide various ways of linking and identifying physical and digital specimens.

Data management platform PlutoF (BiCIKL partner) has been registering DOIs for species datasets for ten years. Currently there are two major registrant types: 1) UNITE Community which uses DOIs to identify and communicate datasets of species. Every DOI based species dataset includes DNA sequences and data on specimens or samples from where the sequences were derived, effectively linking these three data types under the single DOI. Examples of DOI users/consumers are NCBI and ENA who make links between nucleotide sequences and DOIs where they belong. BiCIKL partners GBIF and CoL included the same DOIs into their taxonomic backbone which enables to publish metabarcoding data as taxon occurrences linked to the classification in GBIF. A third user group are researchers and labs

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

who are using those DOI based species datasets for the identification of sequences from samples and specimens; 2) Every registered user or organisation in PlutoF can become a registrant by publishing datasets of taxon occurrences with DOI. The development of the PID system for the specimens and samples is ongoing.
Other communities such as the earth and geological sciences, use IGSN (International Generic Sample Number) which is a handle-based identifier for physical samples. It has been adopted by a growing number of stakeholders worldwide, including national geological surveys, research infrastructure providers, collection curators, and other disciplines that need to refer to physical samples. Bioinformatics and molecular sequence research infrastructures use a mix of different approaches such as compact identifiers (CURIEs) and internal UUIDs (McMurry et. al 2017; Wimalaratne et. al 2018). In addition to physical and digital specimens, there are use cases for unambiguously identifying (for both humans and machines) people (Groom et al. 2022), organisations, and instruments (Plomp 2020).

3. Requirements

3.1. DiSSCo Protected Characteristics

DiSSCo, one of the Research Infrastructures involved in BiCIKL, will be primarily responsible for the specimen level linking. Therefore it is important to highlight the nine characteristics of DiSSCo data management that are essential to protect throughout and ultimately beyond the lifetime of the DiSSCo data infrastructure: (the details of these characteristics are elaborate in the DiSSCo Data Management Plan).

1. Digital Specimen is the core component and the primary digital object type of the DiSSCo architecture
2. Accuracy and authenticity of the digital specimen
3. FAIRness
4. Protection of data (legal regulations and community norms)
5. Preserving readability and retrievability
6. Traceability (provenance) of specimens
7. Annotation history
8. Determinability (status and trends) of digitisation
9. Securability (authentication, authorization, accounting, auditing)

3.2. Technical requirements

These requirements were gathered and synthesised from the DiSSCo PID Consultation, various TDWG conference sessions and other broader global discussions (such as the global consultation on Digital and Extended Specimens).

A PID system should provide the mechanism to create identifiers that are:

- Globally unique
- Persistent

- Discoverable
- Resolvable

Identifiers created by this system:

- Should have appropriate metadata associated with them
- Should never be recycled, reused, or reassigned
- Should never be deleted (see tombstone and PID states note below)
- Can be merged or split if needed or should provide a mechanism to reconcile conflict (for example: a loaned specimen is catalogued by two institutions)

The system should also accommodate the following:

- Lower threshold for curators to register the specimens (and provide tools to import and link the PID to local database)
- Identifiers can be expressed as HTTP(S) URIs
- Support content negotiation for machine representations
- Support discovery APIs
- Offer the ability for institutes to manage their own PID and metadata records
- Allow technology independence: according to the EOSC PID Policy: “Technology independence of PIDs is required to allow for technological change.” This means for instance the PID strings should be generic and independent from technological implementation. But the resolution service relies on specific technical implementations.
- Scalability (billions of digital objects)
- Long term viability (100 years)
- Interoperability with other systems (such as identifiers tracking environmental samples, non DOI based identifiers) and other digital objects (people, institutions, publications)
- Relationships between specimens (such as part-of, derived from)

3.3. Other non-technical requirements

- Have clear documentation
- Have clear usage, creation, and curation policy (example policy from the ePIC consortium). Have training materials, guidelines, and best practices. This will require coordination with various stakeholders and communities (such as CETAF, GBIF, TDWG)
- Transparent, non-profit, long term governance model
- Service continuity and sustainable business model (these are broader general goals that should drive the architectural thinking)
- Clear service specifications (which is abstracted and separated from underlying implementation layer)
- Pre-allocation of identifiers (before the digital specimen is created) (this relates to PID lifecycle)
- Interoperability with existing collection management systems / consider the widely used systems. This points to close collaboration with the CMS development community and user base.

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

- Incorporating and working with existing DOIs (from GBIF, PlutoF, Zenodo)

3.4. Envisaged services

The PID system is a core component for building different tools and services that can be considered as value-added services. Using the PID, associated metadata, and the links can facilitate services such as the following:

- Quality control / error checking: The unique PID string and associated metadata can help with running data quality checks (such as Biodiversity Data Quality from TDWG or different types of FAIR assessment.
- Duplication checking: There should be checks for not assigning different PIDs to the same object
- Citation or impact tracking – along the line of Event data for specimens or Bionomia citation page. PID citation can show impact or usage of the various Digital Objects.
- Reverse lookup (example service: Handle reverse lookup).

4. Components of the PID System

To resolve a PID, an identifier name or PID string **(**example: 10.1038/546033d**)** and a PID resolution system (example: http://hdl.handle.net/, https://doi.org) are needed. Both of these are part of a PID system and have to persist on a technical and organisational level. The PID string needs to resolve to a structured record consisting of a reference and of well-defined attributes (for instance, a Kernel Information Profile) to allow machine and human actions. This section highlights the essential components needed for a PID service to be operational (see Figure 2).

Local Handle Service setup

This is the core component that allows the creation of the identifiers often referred to as “registration” or “minting”. The registration process establishes the reference to the digital object and the additional information. There are also administrative and maintenance functions such as updating of the reference and providing additional information. The resolution of a PID is the process to get the reference and the additional content of the PID. For this we will need LHS and ability to interact with clients and GHS.

PID level metadata and PID Kernel

The PID system can store metadata about the digital object. This is a special type of metadata separate from the content level metadata. Based on the work within the Research Data

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

Alliance, this is often referred to as the PID Kernel Information[3]. The idea is to store a small subset of metadata directly in the PID resolving service to allow fast, machine actionable decisions (Schwardmann and Kálmán 2022). The DOI kernel and DataCite Metadata Kernel also serve similar purposes. The PID-level metadata can provide minimum metadata to describe the object (such as this is a “Digital Specimen” or “Image”) and allow interoperability to combine with other objects

An initial implementation of DiSSCo PID level metadata can be seen in Figure 1. The elements of this PID level metadata records are still under discussion. We also need to ensure that these metadata work with the Minimum Information about a Digital Specimen (MIDS) as MIDS will provide guidelines for different types of elements for digital specimens. However, MIDS is not concerned with the type of object rather what the information element that must be present in each digitization level.

Handle.Net®

Handle Values for: 20.5000.1025/AZW-NVV-KK3

Index	Type	Timestamp	Data
1	pid	2022-09-15 14:29:35Z	https://hdl.handle.net/20.5000.1025/AZW-NVV-KK3
2	pidIssuer	2022-09-15 14:29:35Z	{"id":"https://doi.org/10.22/10.22/2AA-GAA-E29","pidType":"DOI","primaryNameFromPid":"RA Issuing DOI"}
3	digitalObjectType	2022-09-15 14:29:35Z	{"id":"http://hdl.handle.net/21...","pidType":"Handle","primaryNameFromPid":"Digital Specimen"}
5	10320/loc	2022-09-15 14:29:35Z	<locations><location href="https://sandbox.dissco.tech/api/v1/specimens/20.5000.1025/AZW-NVV-KK3" id="0" weight="0"/></locations>
6	issueDate	2022-09-15 14:29:35Z	2022-09-15
8	pidStatus	2022-09-15 14:29:35Z	DRAFT
11	pidKernelMetadataLicense	2022-09-15 14:29:35Z	https://creativecommons.org/publicdomain/zero/1.0/
14	digitalOrPhysical	2022-09-15 14:29:35Z	physical
15	specimenHost	2022-09-19 07:27:56Z	{"id":"https://ror.org/05natt857","pidType":"ROR","primaryNameFromPid":"Needs to be fixed!"}
100	HS_ADMIN	2022-09-15 14:29:35Z	handle=300:0.NA/20.5000.1025; index=200; [create hdl,delete hdl,create derived prefix,delete derived prefix,read val,modify val,del val,add val,modify admin,del admin,add admin,list]
4	digitalObjectSubtype	2022-09-19 07:27:56Z	{"id":"http://hdl.handle.net/21...","pidType":"Handle","primaryNameFromPid":"ZoologyVertebrateSpecimen"}
7	issueNumber	2022-09-19 07:27:56Z	2

Handle Proxy Server Documentation
Handle.net Web Site

Please contact hdladmin@cnri.reston.va.us for your handle questions and comments.

Figure 1: *Example PID level metadata for Digital Specimen.*

Types

In order to ensure machine actionability and FAIR, metadata can be registered as Types in a Data Type Registry (the Type is also a FAIR Digital Object). Combination of required and optional elements and Types can be created as Profiles. As of Dec 2021, the work on Type registry is experimental (see Data Type Registry DTR-Test). Most of these discussions are happening in the FDO Forum’s Technical Specification and Implementation Group. Within

[3] From the RDA Recommendation on PID Kernel Information: “PID Kernel Information is information in the form of attributes stored within the PID record, i.e., information stored at a global or local PID registry and accessible by a resolver. PID Kernel Information supports smart programmatic decisions that can be accomplished through inspection of the PID record alone. PID Kernel Information profiles are registered schemas for PID records. PID records may be created according to specific profiles and checked for conformance against them. In other words, PID records are concrete instantiations of profiles, comparable to how we consider objects as instantiations of classes in object-oriented Programming.” After the RDA, further refinement work on PID Kernel

DiSSCo, we have identified several Digital Objects besides Digital Specimens that will need to be persistently identified: Digital Collection, Collection Descriptions, Images, Annotations, Organisations, Workflows. At the moment, we are envisioning that Type registration should not be mandatory (at least the digitisation and PID registration workflow should not be dependent on a Type registry).

Metadata Profiles

Metadata profiles[4] are important aspects for discoverability and for covering a range of applications that do not require deep access to digital specimens. For example the lower level elements in MIDS (level 0 and 1) for Digital Specimens. There could be other types of profile for other digital objects such as collections, measuring instruments.

API/User Interface

The registration framework needs interfaces: web interface, API, command line tools if needed. Admin interface and APIs also need administration functionalities, for instance updating membership information of the RA or Registrants (consortium member).

Authentication and authorization infrastructure

Authentication and Authorization Infrastructure (AAI) provides the mechanism for access control and authorization. The PID system needs a mechanism to control creation and maintenance of the identifiers. The AAI also needs to allow both API and web interface level interactions and should align with the community standards, requirements and recommendations for AAI (such as OAuth, openID, SAML). Various policy decisions here need to be accommodated as well (for example, should journals and publishers be allowed to register identifiers? if small institutions cannot register their own identifiers, can the RA or another institution do this on their behalf?)

Policy, namespaces and governance

The RA has to have the capability to manage namespaces (top level, second level etc.). This should come with governance and policy documents. These non-technical aspects will drive

[4] What are the differences between PID Kernel Information, Types and Metadata Profiles? Put simply they are all attributes of FAIR Digital Objects. Kernel Information is an attribute stored within the PID record. Each community can have their own Kernel Information. Types are specifically being discussed in the FDO Forum within the context of machine actionability. Types can be thought of as elements that are characterising the digital objects for machines to perform operations. Metadata profiles can be thought of as domain or object specific attributes – metadata for a botanical specimen might be different from a paleontological specimen (the Kernel and Types data might be common elements).

various technical requirements (for example, policy for how to handle batch registration, how to handle backup, pid string requirements, how to outsource operations if needed).

Batch/bulk registration/Message brokering

Various workflows and services, and users will interact with the PID system. The system needs the ability to submit a large number of requests and also support reservation mechanisms (see lifecycle). Asynchronous processing might help with scalability however synchronous processing is preferred to be able to return results back immediately (see the Current Status section for the DiSSCo implementation example). A message broker such as RabbitMQ/Kafka can be used if synchronous processing is not feasible (see example implementation using RabbitMQ for climate data).

Landing page and Signposting

Landing pages are primarily for human agents to provide structured metadata, listing identifiers and link to the content/bit stream. But oftentimes this is not optimised for machines. Signposting is a lightweight interoperable way to provide structured information in a uniform machine-friendly way (see Tombstone page specifications section below for another type of landing page). There are two categories of users to consider: a reader follows the cited PID link. This should be presented as a user-friendly web page. For all other advanced user categories (includes non-human agents) core metadata, partial information can be presented in a JSON or other serialisation format.[5] We acknowledge that there are already different types of landing pages out there for specimen and specimen related records. We will need to consider the current practices before suggesting new standards and best practices. Questions to consider: Who will maintain these landing pages? Where will these pages be stored? Can small institutions maintain and provide data for these pages in a standard way? What will be included in these pages? What is the role of the data aggregators and repositories in this context? Some of these questions need wider consultations and the decisions need to be made by the RA that will be in charge of the PID infrastructure. For DiSSCo the preference is to provide the landing pages through the DiSSCo digital specimen infrastructure. The DiSSCo RI can reliably provide these, while collection providers in the infrastructure may not have the capabilities to do so. Also, collection providers may not have the capacity to store additional data that can be added through community curation and enrichment services operating on the digital specimen objects.

Since PIDs will never be deleted but the object that it identifies may no longer exist, the landing page should sometimes resolve to a tombstone page which is a special type of landing page describing the item that has been removed or deprecated. Tombstone pages should, according to a survey at Pidapalooza (2018), at least include metadata to:

- disambiguate the registrant (who registered the PID), if possible with contact information
- confirm you have found the correct item which for a specimen requires metadata like object name, locality, collection, collector

Respondents to the survey also indicated to want:

[5] Examples: GBIF: https://doi.org/10.15468/5psmq8 Uniprot: https://www.uniprot.org/uniprot/A0AV18

- explanation of what a tombstone page is and confirmation that the PID is valid, just no longer bound to the object
- reason why the PID is no longer bound to the object
- what the user can do next (e.g. link to new DOI) or way to learn more
- perhaps a link to an archived version, like web archive search

Datacite also provides best practices for tombstone pages.

Landing pages should include a proposed bibliographic citation for the DOI. The format for this needs to be further discussed by the community. It could be only the DOI or something more detailed like what is done in IGSN, e.g. Chan, R.A. "Sample AU1243". A digital catalogue record of a physical sample managed by Geoscience Australia. Accessed 15 October 2022. igsn:AU1243., though since the DOI is resolvable that is not really needed. When including all identifiers it could become fairly complex, e.g. *Theodor Hesselberg (https://www.wikidata.org/wiki/Q3429562) et al. (1907-2022): preserved specimen, urn:catalog:O:V:422965, https://doi.org/10.1.10/ABCD.EFGH (SHA1 76a4acaf31b815aa2c41cc2a2176b11fa9edf00a), in University of Oslo (https://ror.org/01xtthb56)*. It seems enough to include only the DOI and a hash or version number for the data object.

PID States/Lifecycle management

PIDs are essential elements of FAIR data workflows. Similar to data lifecycle, PIDs and associated metadata need to accommodate changes based on the workflows. For instance, a client can request and reserve a large number of identifiers for later use. As the status of a digital object changes, PID metadata should reflect such status as well.

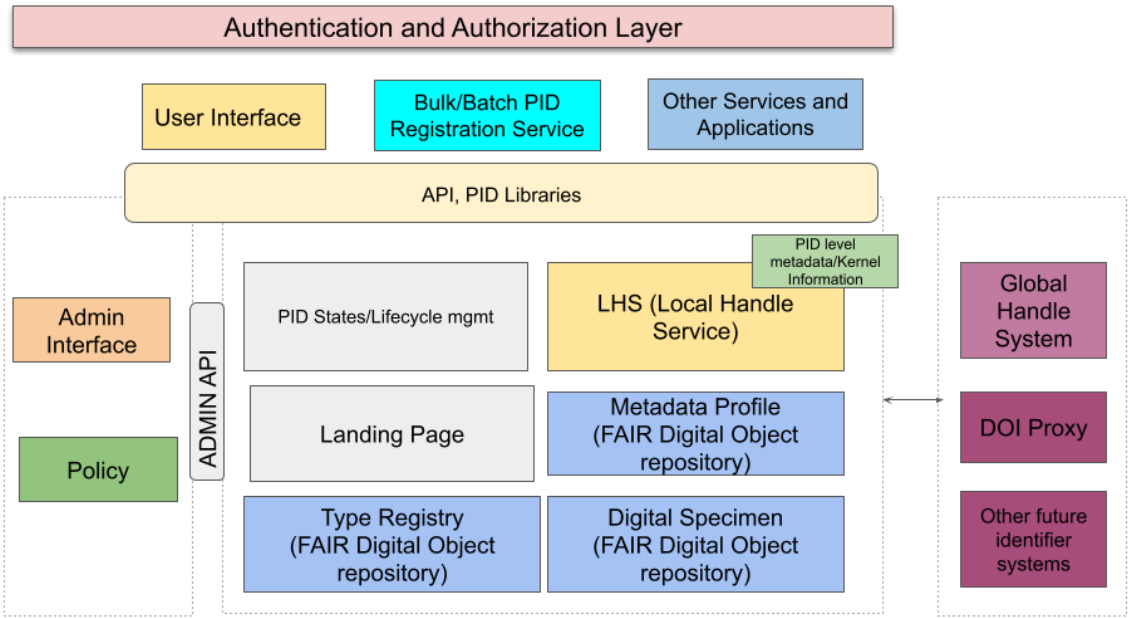

Figure 2: *PID system components.*

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

5. DiSSCo Implementation

A preliminary setup of the PID system has been implemented within the DiSSCo Prepare project. This section describes the setup (as of October 2022) used within the pilot for the DiSSCo core infrastructure. Based on implementation experience and discussions with BiCiKL, DiSSCo, and CNRI team members, we came to the solution shown in Figure 3. There are three principal components involved in the generation and resolution of the Handles: the Local Handle Server, the Handle API and the Processing Service.

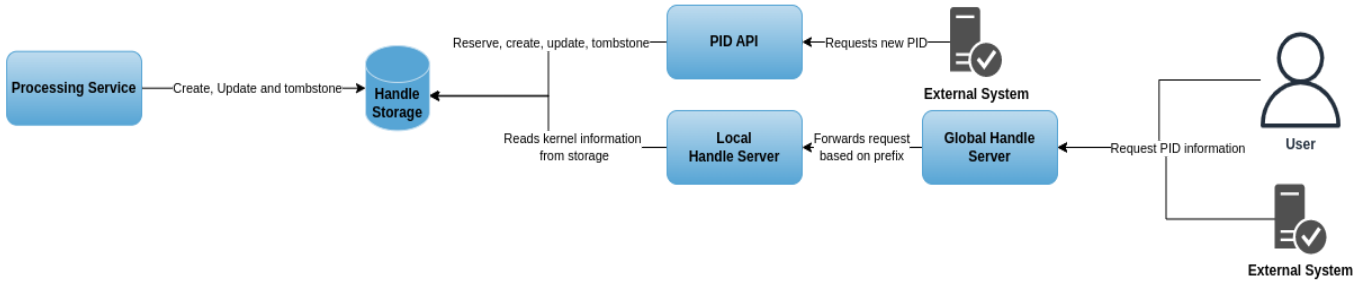

Figure 3: *Current PID system implementation within DiSSCo*

The Local Handle Server

The Local Handle Server is a core part of the Handle resolution system. After a prefix is created, the Global Handle Server redirects the request to the local server based on the Handle prefix. This is essential for the global server to find the local server. CNRI provides a packaged Handle server which can be installed and used for all operations. However, during development, we came across a few issues, especially with the creation of large amounts of PIDs. With the help of CNRI, we implemented a solution by switching the data storage solution behind the local Handle server to an external database.[6] This way, we could shift the create, update, delete functionality to our own services (see Custom Handle API below) while keeping the Local Handle Server for PID resolution. This server makes sure the PIDs are findable through the Global Handle Registry and the PID kernel information is exposed and can be resolved.

Custom Handle API

Besides the issues regarding the creation of large amounts of PIDs, there were further advantages to building our own custom API for managing handles. We can now enforce PID level metadata (Kernel profile) discussed earlier. DiSSCo-registered Handles now follow a specific structure and adhere to our data model. We can also enforce the desired PID life cycle based on digital specimen workflow. Additionally, developing an in-house API allows us to enforce our own Authentication and Authorization Infrastructure, rather than relying on the default private-public key pair distributed by CNRI.

[6] Currently this is a AWS managed Postgresql database. We saw some improvements there which we discussed with CNRI. In the next local handle server version, version 10, MongoDB will be supported as a storage solution to which we might switch.

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

The API interfaces directly with the Handle Server's storage facilities, allowing for faster and more flexible operations. For instance, because the API directly communicates with the Handle database, we are able to implement batch operations (sending several operations together instead of one after another).

The current API is a Java-based RESTful API built using Spring Boot[7]. Endpoints are established for the following functions:

- POST Operations
  - Create a new handle
  - Reserve a batch of handles
  - Update a handle record
- GET Operations
  - Resolve handle record
  - Get list of handles
  - Get subset of handles, select based on pidStatus (this can be further extended to other data fields if desired)
- PUT Operations
  - Merge handle records
  - Split handle record
- DELETE Operations

Processing Service[8]

While most of the above actions will manage a single or a batch of PIDs, we also need to handle a large stream of objects during data ingestion. For these objects such as digital specimens, digital media, and annotations, we need to create PID records. As this will represent a large number of PIDs to be minted, we need a fast and resilient way to create, update, or tombstone these objects.

For example, when a digital specimen is ingested into DiSSCo from a Collection Management System (CMS), the Processing Service verifies whether it is a new specimen or not. Because DiSSCo Digital Specimen PIDs are not yet part of partner CMS infrastructures, the Processing Service needs to query the unique physical specimen identifier used by the CMS. If the CMS uses a CETAF Identifier, the Processing Service will query that.

If the CMS only uses its own internal identifier system, there is no guarantee those identifiers are unique within DiSSCo. As a result, a new identifier needs to be created by the Processing Service and queried. One option here for this new identifier is a combination of the provided CMS identifier, the organisation identifier (i.e. ROR), and possibly the collection number (this collection number is not part of the current DiSSCo implementation, subject to ongoing discussion). This uniqueness is important during the digitisation pipeline and MIDS-0 (which is a pre-level where records are generated prior to any formal digitisation) to ensure that a new unique digital specimen identifier is being connected to the unique physical identifier. This challenge of identifying new specimens within the DiSSCo infrastructure further highlights the need for a consistent, globally recognized PID system for collections.

[7] Github Repository

[8] See for the code of the three different types (digital specimen, digital media and annotations) of processing service https://github.com/DiSSCo the project prefixed with 'dissco-core'

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

Next the Processing Service conducts a lookup using either the globally unique physical identifier or the ID generated by the Processing Service. If nothing is returned, we can be ensured that the incoming digital specimen is new. Then the Processing Service will generate a new Kernel PID Record (as in Figure 1) and create a new PID. After a new PID was successfully created, the specimen data will be inserted in the database and indexing service. A notification will also be sent to external systems. This way the CMS could collect the created Digital Specimen PID if needed.

If the digital specimen already exists, the Processing Service identifies what information, if any, is different between the existing and incoming specimen (this is similar to performing a diff operation in Linux). The existing record is updated with the new information and increments the record version number. The Processing Service also checks changed fields and updates the Kernel if it has changed. If the Kernel is updated, the issueNumber is incremented by one and the issueDate is set to the current date.

In the section below, we describe different PID states that reflect the data lifecycle. For example, an object could be deprecated – either it was misplaced or lost. It could also be that the object was created by a user, such as an annotation, and the user wishes to remove the object. For these cases, the system will interact with the kernel information and change the status and if provided, fill in a tombstoneText to explain the tombstoning of the record.

Example PID level metadata (PID kernel) for Digital Specimens

Table 1 shows our current implementation (as of Oct 2022) of the PID kernel metadata for Digital Specimens. For the updated version of this metadata please visit the DiSSCo Digital Specimen PID github page.

Table 1: *Sample PID level (kernel) metadata.*

Index	attribute	example
100	HS_ADMIN	handle=20.5000.1025/2FAH-GB4Y; index=300; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]
1	pid	https://doi.org/10.22/2FA-HGB-4Y3
2	pidIssuer	"{ "pid": "https://doi.org/10.22/10.22/2AA-GAA-E29", "pidType": "DOI", "primaryNameFromPid": "RA issuing the DOI", "registrationAgencyDoiName": "10.22/2AA-GAA-E29" }"

3	digitalObjectType	"{ "pid": "https://hdl.handle.net/21...", "pidType": "Handle", "primaryNameFromPid": "Digital Specimen", }"
4	digitalObjectSubtype	"{ "pid": "https://hdl.handle.net/21...", "pidType": "Handle", "primaryNameFromPid": "Botany Specimen", }"
5	10320/loc	<locations> <location id="0" href="https://sandbox.dissco.tech/api/v1/specimen/test/bff3e176-7ace-45f0-b40e-c3d8dd495de1" weight="0" /> <location id="1" href="http://www1.example.com/merged_specimen_X" weight="1" status="obsolete" /> <location id="2" href="http://www1.example.com/merged_specimen_Y" weight="1" status="obsolete" /> </locations>
6	issueDate	2022-11-24
7	issueNumber	2
8	pidStatus	ACTIVE
9	tombstoneText	The specimen was merged with specimen Y.
10	tombstonePids	{[{"pid": "https://doi.org/10.22/...", "primaryNameFromPid"": "Specimen X"}, {"pid": "https://doi.org/10.22/...", "primaryNameFromPid"": "Specimen Y"}] }
11	pidKernelMetadataLicense	https://creativecommons.org/publicdomain/zero/1.0/
12	referentDoiName	10.22/2FA-HGB-4Y3

13	referent	{
		"primaryReferentType":"materialsample",
		"referentMaterialsample":{
		"materialsampleName":{
		"value":"Tulostoma brumale Pers.",
		"type":"Name",
		"primaryLanguage":"en"
		},
		"identifier":[
		{"nonUriValue":"WAG0383181@BRAHMS", "type": "ProprietaryIdentifier", "userDefinedType": "local identifier in Brahms collection management system"},
		{"uri":

		"https://bioportal.naturalis.nl/specimen/WAG0383181",
		"type":"URI"}
		],
		"structuralType":"physical",
		"mode":"tangible",
		"character":"other",
		"type":"Specimen",
		"principalAgent":{
		"name":{
		"value":"DiSSCo",
		"type":"PrincipalName"
		},
		"identifier":[
		{"value":""https://doi.org/10.22/10.22/CDA-HBA-EA6",
		"type":"DOI"}
		],
		"role":"Digital Specimen provider"
		}
		}
		}

14	digitalOrPhysical	physical

15	specimenHost	{
		"pid": "https://ror.org/0566bfb96",
		"pidType": "ROR",
		"primaryNameFromPid": "Naturalis Biodiversity Center",
		}

16	inCollectionFacility	{
		"pid": "https://doi.org/10.22/10.22/CHH-A8A-CC3",
		"pidType": "DOI",
		"primaryNameFromPid": "Botanical Collection",
		}

17	objectType	{
		"Organism material": {
		"objectType": "herbarium sheet"
		}
		}

18	preservedOrLiving	preserved

black = kernel metadata used for all digital objects
blue = extra metadata for DOI
red = extra metadata for Digital Specimen
green = extra metadata for BotanySpecimen

6. PID LifeCycle

As mentioned above, PID lifecycles, similar to data life cycles, are important concepts to accommodate different objects and workflow processes. These life cycles will be recorded as pidStatus in the PID record (see Table 1 above). For a similar implementation, see DataCite

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

DOI states. The lifecycle states we currently envisage are listed in Table 2.

Table 2: *PID states.*

PID State	Description	Notes
DRAFT	DRAFT status means the PID record is not published and findable in the Digital Specimen repository.	DRAFT status can be used for digital objects that are work in progress. Depending on the implemented policy, DRAFT PIDs can be deleted if needed. Also DRAFT PIDs can be created without metadata or resource location. The Handle system does not have the concept of DRAFT, therefore internal check and pidStatus will be used as a metadata element to control access and findability. Another use case of DRAFT could be for records that need to be embargoed for a certain period (for licensing or legal reasons).
ACTIVE	PIDs are registered in the global system and indexed in the Digital Specimen repository.	ACTIVE here denotes that objects are findable and usable for all purposes.

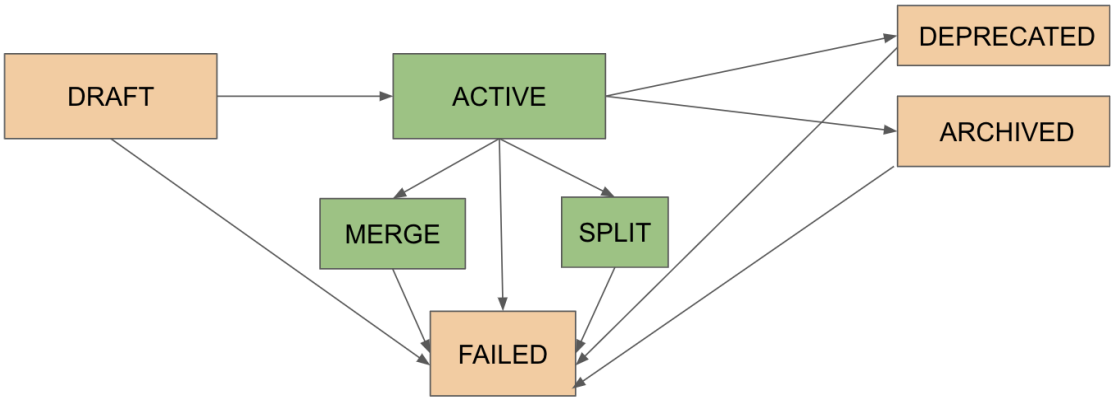

Figure 4: *PID lifecycle.*

There are other aspects of the specimen data curation that we need to consider, capture, and store for operational and historical purposes. For example, merging, splitting, or destruction of a physical specimen. These are not the properties of the identifier but more of a metadata element for the digital specimen object; however, the pidStatus needs to capture the state in the data lifecycle. Table 3 shows a few examples. These examples are in the early stage of conceptualization. We have tested that the current PID system design can allow different states. The specific workflow and metadata generation need to be worked out in further

ARPHA Preprints *Author-formatted document posted on 30/05/2023.* DOI: https://doi.org/10.3897/arphapreprints.e107168

details.

Table 3: *Metadata for operational purposes.*

Metadata term	Description	Notes
ARCHIVED	A resource that has been revoked.	For legal reasons, certain specimens can be deaccessioned from a collection. The record still should be kept for historical reasons.
DEPRECATED	A digital object that is not fully operational or active.	The reason for the deprecation can be use case specific or dedicated by policy. We need to capture the reason for depreciation as well. If a resource (physical or digital) has been removed, lost or stolen, the metadata should reflect the status. In case of deleted resources, the PID should resolve to a Tombstone page (McMurry et al. 2017).
FAILED	Optional: Unexpected error has occurred during assigning the PID to a resource.	This could be handled via a roll back mechanism so this state might not be needed in most cases.
MERGED	When two or more identifiers merge, a new identifier should be designated and information about the legacy identifiers should be present and should resolve to the new identifier.	We need to see what are allowable use cases for this. These are also policy related decisions.[9]
SPLIT	If an identifier is split into 2 or more, new identifiers should be assigned. A connection and relationship should be maintained with the original object.	Splitting of the physical specimen can also occur. This use case can vary across organisations.

7. Inclusion of existing identifier systems

Digital Specimens in the FAIR data ecosystem will need to work and integrate with other types of digital objects. For example, digital specimen images or collection descriptions.

[9] Example of a record that has been demerged: https://www.uniprot.org/uniprotkb/P29358/history

These objects can be cited in a paper or used within a workflow system. Objects that already have an authoritative identifier like ROR (for research organisations) or ORCID (for researchers) do not need a DOI but can have a Handle (representing a Digital Object describing the entity). Then this new digital object can include the authoritative identifier in the metadata plus additional metadata that is not stored in the authoritative identifier.[10] There could be technical and operational implications for including these elements within the metadata as these links need to be indexed. A tool such as ElasticSearch could be used to create indexes that contain links between objects. Depending on the use case, It may be beneficial to give each related and linked object its identifiers. For example, a list of all the specimen media objects can have their own PID. To start with, we need at least DOIs for collection descriptions, digital specimens and their media objects. To allow for collection inventories based on species or storage unit lists, we may also need identifiers for these.

8. Conclusion

A pan-European PID system is a key component for realising the vision of DiSSCo and BiCIKL – a FAIR data ecosystem that connects specimen and specimen related data with wider data classes. The Digital Specimens as FAIR Digital Objects (with PID at the core) will provide the building blocks for a wide array of services that can advance data driven and multidisciplinary research. Together with FAIR principles and other community standards, this architectural design of a PID system for Digital Specimens reflects wide collaboration, and a mix of technical, social, and organisational insight. This design document will serve as the foundation for further development.

9. Acknowledgements

We would like to thank everyone who reviewed and commented or otherwise contributed to this document. In particular we would like to thank: Sam Leeflang, Jonathan Clark, Lyubomir Penev and Tim Robertson for their valuable comments.

10. References

Boschert and Dikow 2022

Hardisty, A.R., Ellwood, E.R., Nelson, G., Zimkus, B., Buschbom, J., Addink, W., Rabeler, R.K., Bates, J., Bentley, A., Fortes, J.A. and Hansen, S., 2022. Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. BioScience, 72(10), pp.978-987. https://doi.org/10.1093/biosci/biac060

Hardisty, A., Addink, W., Glöckler, F., Güntsch, A., Islam, S. and Weiland, C., 2021. A choice of persistent identifier schemes for the Distributed System of Scientific Collections (DiSSCo). Research Ideas and Outcomes, 7, p.e67379. https://doi.org/10.3897/rio.7.e67379

[10] For example, an EC PIC number or a visiting address for an organisation is not in the ROR metadata.

Groom, Q., Bräuchler, C., Cubey, R., Dillen, M., Huybrechts, P., Kearney, N., Klazenga, N., Leachman, S., Paul, D.L., Rogers, H. and Santos, J., 2022. The disambiguation of people names in biological collections. Biodiversity Data Journal, 10, p.e86089. https://doi.org/10.3897/BDJ.10.e86089

Güntsch, A., Hyam, R., Hagedorn, G., Chagnoux, S., Röpert, D., Casino, A., Droege, G., Glöckler, F., Gödderz, K., Groom, Q. and Hoffmann, J., 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, R., Conlin, T., Deck, J., Stucky, B.J. and Cellinese, N., 2014. The trouble with triplets in biodiversity informatics: a data-driven case against current identifier practices. *PloS one*, *9*(12), p.e114069. DOI: https://doi.org/10.1371/journal.pone.0114069

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

McMurry, J.A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., Courtot, M., Deck, J., Dumontier, M., Fellows, D.K. and Gonzalez-Beltran, A., 2017. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS biology*, *15*(6), p.e2001414. DOI: https://doi.org/10.1371/journal.pbio.2001414

Plomp, E., 2020. Going digital: persistent identifiers for research samples, resources and instruments. Data Science Journal, 19(1). http://doi.org/10.5334/dsj-2020-046

Sabaj, M.H., 2020. Codes for natural history collections in ichthyology and herpetology. *Copeia*, *108*(3), pp.593-669 DOI: https://doi.org/10.1643/ASIHCODONS2020

Schwardmann, U., 2020. Digital Objects–FAIR Digital Objects: Which Services Are Required?. *Data Science Journal*, *19*(1) DOI: http://doi.org/10.5334/dsj-2020-015

Schwardmann U, Kálmán T, 2022. Two Examples on How FDO Types can Support Machine and Human Readability. Research Ideas and Outcomes 8: e96014. DOI: https://doi.org/10.3897/rio.8.e96014

Wimalaratne, S.M., Juty, N., Kunze, J., Janée, G., McMurry, J.A., Beard, N., Jimenez, R., Grethe, J.S., Hermjakob, H., Martone, M.E. and Clark, T., 2018. Uniform resolution of compact identifiers for biomedical data. *Scientific data*, *5*(1), pp.1-8. DOI: https://dx.doi.org/10.1038%2Fsdata.2018.29