H2020-INFRADEV-2019-2
Grant Agreement No 871043

DiSSCo Prepare report D6.1 Harmonization and migration plan for the integration of CMSs into the coherent DiSSCo Research Infrastructure - MfN WP6/T6.1

Work package lead: Claus Weiland (SGN)

Task lead: Falko Glöckler (MfN)

Version 1.1 [27.07.2022]

https://doi.org/10.34960/366d-sf49

Authors: Falko Glöckler, Julia Pim Reis, Sabine von Mering, Mareike Petersen, Claus Weiland, Mathias Dillen, Sam Leeflang, Elspeth Haston, Wouter Addink, David Fichtmüller

Abstract

DiSSCo Prepare Deliverable D6.1 describes the concepts for the integration of collection management systems (CMS) into the DiSSCo Research Infrastructure (RI), including recommendations for APIs guidelines. By identifying the challenges of this integration, the basic requirements for the harmonization of CMSs and subsequent interoperability with DiSSCo's central Digital Specimen services can be identified and addressed. One of the main challenges is the synchronization of specimen data in the DiSSCo RI with data in local CMSs of collection-holding institutions, as these represent a high diversity of data models, software frameworks and API capabilities. Furthermore, different workflows and use cases from multiple disciplines need to be accommodated, while the technological diversity of CMSs necessitates content-wise abstraction in a sustainable implementation. With the help of results from an Event Storming workshop with a mixed group of curators, collection managers, CMS users and vendors, the most important event types in CMSs have been identified, aggregated and classified to allow for a harmonized, formal description of events so they can be used for standardized communication between the CMS and the DiSSCo RI. The DiSSCo RI will provide an abstraction layer through FAIR Digital Objects (FDO; see Islam et al. 2020) for the heterogeneous data. The harmonization of events and the abstraction layer together provide a solid foundation for data aggregation and interoperability. The Digital Objects are serialized as JSON and DiSSCo promotes the use of open standards and open software. Therefore, we suggest API guidelines derived from the existing specifications of JSON:API, OpenAPI and CloudEvents.

Keywords

DiSSCo Research Infrastructure, Collection Management System, Interoperability, API Guidelines, FAIR, Digital Object Architecture, Data Harmonization, Digital Specimen, Natural History Collections

Index

1. Introduction

The Distributed System of Scientific Collections (DiSSCo) Research Infrastructure (RI) is working towards a digital unification of all European (but also global) natural science assets, under common curation, access policies and practices in order to make the data more Findable, Accessible, Interoperable and Reusable (FAIR). Within DiSSCo's preparatory phase project (DiSSCo Prepare) the work-package Technical Architecture & Services Provision (WP6) aims at bringing the design of the DiSSCo technical architecture to the required maturity by providing a comprehensive implementation plan. A harmonized connection and integration between the DiSSCo RI and collection management systems (CMS) is key to the interoperability of the digital representations of collection objects curated as Digital Extended Specimens.

CMSs are electronic collection catalogs used to document all events related to a certain curated object in a collection, by storing (meta-)data about the specimens such as their provenance (e.g. collection location, date, collectors), their scientific identifications, their physical state and their location in the collection. As such they are often a combination of an asset management system and a repository for scientific information. In addition, the systems sometimes also include features to organize transactions such as acquisition, loan, and deacquisition. As the digitization of collections proceeds, CMSs have also become important by providing a home for the verbatim data digitized from analog collection catalogs, labels, filing cards, or arrival books.

Challenges

The integration of CMSs into the DiSSCo RI is facing some challenges that need to be met, not only the high diversity of systems used in collection holding institutions, but also the difficulty in keeping the physical and digital collections synchronized and connected as their data can be amended independently and local identifiers for the physical objects tend to change over time. Specimens may be physically handled during conservation treatments, scientific research, curatorial events, rehousing or shipment for loans. And even without changing conditions in the storage environment, the physical specimens may still deteriorate over time to varying degrees. Ideally, these changes are documented in the CMS in order to keep track of the collection's inventory, the objects' provenance, availability and their general condition. Specimen data may also be enriched through research results such as updated determinations, measurements, references to literature and other metadata. These results

are usually added or linked to the records in the electronic catalogs similar to the methods used in the analog documents.
Digital representations (Digital Specimens; Hardisty 2020, Hardisty 2018) of the real-world objects are also subject to changes. A digital specimen can be used for research without making direct use of the physical one. For example, digital images at sufficient resolutions could be used to identify a species. Thus, one could record a new determination of the physically preserved object, but it would strictly speaking be a determination of the digital specimen. An efficient propagation mechanism will be needed to connect this new information back to the physical specimen as the metadata of the determination would initially reside only with the digital object. This would also be true for any other additional information derived from both, the digital specimens alone and the research on the digital specimens. Therefore, one could state that in a virtual research context such as the DiSSCo RI, the digital objects may potentially evolve as distinct digital twins more or less independently from the physical objects, despite the fact that they are closely related to each other and still act as proxies in research. This becomes obvious in cases when the physical specimen is destroyed or lost, but the digital specimen is subject to continuous research. Given this as a premise, the following challenges can be identified:

Challenge 1 - Syncing basic information

Digital specimen records are often initially created with the label information of the physical specimens (e.g. metadata of the collecting event, verbatim information from the labels, descriptions and photographs of the physical objects, the curatorial information relating to where the specimen can be located and the specimens catalog numbers and/or persistent identifiers like the CETAF Stable Identifier (Güntsch et al. 2017) etc.). But this digital record can be continuously added to throughout the whole lifetime of a digital specimen, as long as knowledge of the physical object is being gained (Lendemer et al. 2019). Furthermore, operations on the digital specimens may be used as a proxy for the physical objects. For example, a loan of the physical specimen might be requested through its digital representation which would have implications for the (temporary) availability of the physical specimen once the loan request has been actioned. The physical filing location of the specimen within the collection may only be present in the digital specimen and, if the physical specimen is moved, this information would need to be added to or updated in the digital specimen record. Furthermore, some research outcomes concluded from the digital specimens might gain more impact if they were additionally verified with the help of evidence from the physical specimen. It is important to distinguish the references (persistent identifiers) of the digital and physical objects while citing the research material used in the

scientific work. For example, it must be very clear that the CETAF Stable Identifiers identify the physical specimen, but as these cannot be transferred via the internet, the CETAF ID redirects to the digital data as a surrogate (Güntsch et al. 2017).
Hence, maintaining the synchronization of data on physical and digital specimens is one of the key challenges in the integration of CMSs into the DiSSCo RI to allow for an appropriate usage and maximum benefit of the resources available.

Challenge 2 - Work processes

The digital transformation of science relies on interweaving the physical with the digital processes. This is highly relevant regarding the documentation of research methods, workflows and results as well as treatments, provenance and transactions of physical and digital objects. Digital workflows should help keep track of all this, ideally in a standardized way. More sophisticated CMSs might try to accommodate features to facilitate the documentation of the daily work with the physical specimens instead of only holding the inventory of a collection. Thus, the boundaries between the physical and digital workflows become blurred as both are digitally recorded. Basically, this is an advantage in terms of leveraging the use of digital surrogates, e.g. using digital images (if appropriate) instead of overstraining the physical objects through extensive handling. However, there is the challenge to avoid disadvantages such as ambiguous documentation of research methods and citation of objects in the DiSSCo RI.

Challenge 3 - Diversity of Systems

There is a huge variety of CMSs available in Europe and world-wide, each designed with different focus on specific disciplines or purposes, and/or differing in their functionality, technology and data models. This diversity across different institutions is a challenge for their integration in global infrastructures like the DiSSCo (compare Dillen et al. 2019), because the harmonization on a technical level must consider not only the content-related differences, but also the different feature sets and capabilities to handle similar (but not identical) workflows in a harmonized manner with interfaces (e.g. APIs) across communication with other systems and services.
Besides the technical aspects of the different systems, the challenge of implementing diverse systems into the DiSSCo RI has also a social component. The end-users of CMSs need guidance (e.g. in the user interfaces) to allow for a wider view across the boundary of the local system, so they can take well-advised steps in their daily work with the integrated system.
In order to allow the technical and social challenge to be addressed, the harmonization and

integration of different CMSs need comprehensive abstraction layers that would cover the use cases most important for DiSSCo RI.

Challenge 4 - Resources

As for all pieces of software in any user scenario, long-term maintenance and active further development are also key to the CMS's successful harmonization and implementation into the DiSSCo RI. However, even for vendors and organizations with well-organized and economically sustainable business models, estimated costs for the initial implementation and mid- to long-term maintenance are essential for any decision towards DiSSCo technical readiness.

The costs would be on an individual institutional level given the diversity of systems, technology used and specific needs of their end users.

To allow for an individual estimate of costs per system the vendors and/or developers would need a blueprint to serve as a guideline or checklist of criteria to be implemented. Thus, the efforts necessary for the technical readiness can be calculated as a predictable budget based on the individual parameters (compare Petersen et al. 2022).

Challenge 5 - Prioritization

As each CMS has a slightly different focus (e.g. regarding collection type, specimen type, scientific discipline) the CMSs might differ in their preference of certain features and workflows they would like to implement for the DiSSCo RI. This would mirror the diversity of systems in the support of DiSSCo's service APIs which might cause conflicts by serving all relevant stakeholders. Thus, the challenge for harmonizing CMSs for integration into the DiSSCo RI is to create an abstraction layer between the systems and services. The specification of such an abstraction layer needs to be flexible and generic to accommodate a high number of use cases without creating too many individual solutions (e.g. per discipline).

Challenge 6 - Evolving systems

Vendors of CMSs who will take the initial step to follow a DiSSCo blueprint for implementation, would still be facing the challenge of modifying their systems for both, new features requested by their users, and updates through external technology changes. As this would be true for any kind of software (consequently also the DiSSCo services), the key challenge in the context of technical readiness is that all systems may evolve independently. Thus, it would need a communication and coordination mechanism to avoid incompatibility and to check and certify DiSSCo compliance.

2. Methods

Specification for Abstraction Layers

Each CMS may have different API endpoints (if any) to communicate with the DiSSCo RI. This means that each CMS would have individual specifications on how its API should interact with DiSSCo services, and all of these individual approaches would need to be documented as well. Clients, applications and services (like those related to / developed by the DISSCo RI) planning to interact with several CMSs would have to invest huge efforts to implement and update each of the CMS-specific methods. Furthermore, communication with the CMSs and/or processing the data would likely be unreliable as soon as the specific APIs change, because it would be hard to keep track of the changes. Therefore, a common specification or guideline for designing an API which supports interacting with the DiSSCo RI can help harmonize the diverse ways of communication via these programming interfaces.

The DINA[1] consortium (Glöckler et al. 2020) agreed on using API guidelines (https://github.com/DINA-Web/guidelines/blob/master/DINA-Web-API-Guidelines.md) in order to establish a uniform communication layer between different web-based software modules and (micro-)services, despite the fact that they were developed separately by different institutions, (potentially) in different programming languages and without implicit dependencies amongst the modules. Additionally, in DINA the modules should be capable of being exchanged by other alternative modules covering the same or extended functionality within the system according to its high-level model of functional components (https://github.com/DINA-Web/dina-model-concepts). This would gain flexibility and adaptability according to the end-user's needs.
In order to achieve this, DINA adopted two community-driven standards in its guidelines : JSON:API specification (https://jsonapi.org/) for building and OpenAPI (formerly known as Swagger) specification (https://www.openapis.org/) for describing and documenting the APIs. These methods can be taken into account for the harmonization of CMSs with DiSSCo as well.

JSON:API is a shared convention for building APIs in JSON format aiming at increasing productivity by uniformity of calls and responses, and at taking advantage of generalized tooling. Therefore, the specification represents an abstraction layer for the requests sent to

[1] DINA ("**DI**gital information system for **NA**tural history data") is a framework for like-minded practitioners of natural history collections to collaborate on the development of distributed, open-source software that empowers and sustains collections management. https://dina-project.net

and the responses received from the API endpoints in order to read, create, and update resources. The JSON:API specification is extensible in order to define new functionality not provided by the base specification. Thus, it allows for flexibility in the abstraction layer even for very specific applications that might hit the boundaries of the base specification.

The **OpenAPI** Specification is a uniform schema designed for describing an API, its endpoints, parameters, responses and formats as well as success and error codes. The documentation comprises a JSON object (in JSON or YAML format) accommodated in a simple text file. Using OpenAPI documents one could follow one of the two traditional approaches: (1) writing the code and generating the API documentation afterwards ("Code-first"), or (2) describing the API and using it as part of the software's blueprints ("Design-first").
Both approaches have their advantages and disadvantages, and both can be interwoven in the continuous integration and continuous deployment (CI/CD) pipelines. Depending on the preferences of the developers, the technical documentation can be automatically generated from code annotations; or the skeleton code for the API can be automatically generated.

For the management of heterogeneous data the **Digital Object Architecture**, DOA (DONA Foundation 2019, Kahn & Wilensky 2006) was evaluated by DiSSCo in order to create a seamless virtual collection of bio/geo specimen data (Lannom et al. 2020, Islam et al.2020). The DOA introduces an abstraction layer for the management of the heterogeneous data of a Digital Specimen. Interoperability of CMSs with the DOA may require support for the Digital Object Interface Protocol, DOIP, which "*[...] specifies a standard way for clients to interact with digital objects (DOs)*." (DONA Foundation, 2018). This can be realized by using and adapting existing components available for and related to the DOA. DiSSCo is piloting an instance of the Digital Object Repository and Registry software *Cordra* as a Digital Specimen Repository and index for the different kinds of natural science objects, their relations and resolvable identifiers (NSIDR = Natural Science Identifier Registry; https://nsidr.org/). A demonstrator was developed in the ICEDIG project (Innovation and consolidation for large scale digitisation of natural heritage, https://icedig.eu/) to test the Digital Specimen concept and will be further developed into a pilot to support development of the Open Digital Specimen specification, openDS (Addink & Hardisty 2020). A demo instance of the NSIDR is available at https://demo.nsidr.org/.
As not all potential use cases and workflows can be accommodated in the pilot, a task group within the DiSSCo Prepare work-package "Technical Architecture & Services Provision" (WP6) organized a workshop with different stakeholders in order to allow for identification, appropriate prioritization and documentation of relevant connection points and potential

dependencies between CMSs and the DiSSCo RI. The workshop was conducted with the help of the Event Storming method (Brandolini, 2013). The results from the workshop will be fed into and provide priorities for the development of the DiSSCo pilot.

In addition to the above-mentioned specifications that may help add abstraction to the technical layers, the content-wise abstraction of data and processes needs to be considered as well. For collection data, common domain-specific standards such as ABCD (https://abcd.tdwg.org/) and Darwin Core (https://dwc.tdwg.org/) would be appropriate to be mapped to openDS. There are existing tools widely used like the BioCASe Provider Software (BPS) (https://www.biocase.org/products/provider_software/) and the Integrated Publishing Toolkit (IPT; https://www.gbif.org/ipt) that are designed for mapping the local CMS databases to these data standards and exposing the standardized data for harvesting by data aggregators like GBIF (https://gbif.org). However, such tools might not be ready to address the needs for DiSSCo, yet, as the harmonization layer between different CMSs and the DiSSCo RI would need mechanisms to inform each other's services as soon as something relevant has been added or changed on either side. Therefore, a common description of such *events* needs to be adopted in order to allow the event *producer* to communicate with the event *consumer*. A simple example for such an event is: The metadata about the physical specimen has been changed in the local CMS and the Digital Specimen should reference the latest version of the metadata (see above "*Challenge 1 - Syncing basic information*"). Thus, the CMS is the event producer and the DiSSCo RI might want to respond to this event as an event consumer. On the other hand, there might be a loan request for a physical specimen in the European Loans and Visiting System (ELViS, https://elvis.dissco.eu/welcome) as part of the DiSSCo RI that needs a response from the collection via the CMS. Thus, the DiSSCo RI would be the event producer and the CMS the event consumer.

In order to allow for the communication between systems about such events, the **CloudEvents** specification (https://github.com/cloudevents/spec/) can be adopted. The CloudEvents project (https://cloudevents.io) is "[...] *working to formalize the specification based on design goals which focus on interoperability between systems which generate and respond to events.*" (CloudEvents 2022). The basic principle is that an event contains two types of information: event data and context. The context is the metadata about the event itself, which usually comprises the version of specification (`specversion`), the type of event (`type`), the `subject` of the event, the `time` of occurrence, the event identifier (*id*) and the mime type of the data transferred (`datacontenttype`). The event's `data` is the actual,

domain-specific payload (e.g. in the context of CMSs and DISSCo: a new or changed specimen record from the CMS or an array of identifiers pointing to the new or changed records) that would be transferred for the specific consumption by the counterpart if the particular event type is of interest.
An example for an emitted event described in JSON format would be:

```
{
    "specversion": "1.0",
    "type": "org.dissco.event.object.created",
    "source": "https://collection.myinstitution.com/dissco/event"
    "subject": "item 123",
    "id": "A234-1234",
    "time": "2022-02-06T17:31:00Z",
    "datacontenttype": "text/json",
    "data": "{ // put the payload here }"
}
```

In the example, an event of adding a new collection object to the CMS is described. Therefore, the event type “object created” could be defined as reverse domain name notation to indicate the DiSSCo context as “`org.dissco.event.object.created`”. The `source` would be an API endpoint of the CMS implementation in the respective collection holding institution and the `subject` would be the local (or even better globally resolvable) identifier (e.g. “`item 123`”) of the affected object. This must not be confused by the identifier (“`A234-1234`”) of the event, which can be produced e.g by the audit log of the CMS.

In order to use the CloudEvents standard the identification and descriptions of event types is key to a successful adoption of this harmonization and abstraction layer. Therefore, the Event Storming method was used to start the discussion on the most relevant events in the communication between CMSs and the DiSSCo RI.

Event Storming Method

Event Storming is a lightweight technique created by Alberto Brandolini (https://eventstorming.com) for rapidly modeling a process that consists of an unlimited number of events along a timeline. For this purpose, anything that occurs in the process is described as a domain event triggered by an agent ("*actor*") doing a certain activity ("*command*"). An agent could be a human being, but also a machine or software triggering an event. Based on each event one or many reactions ("*responses*") can be defined, which may be described as domain events as well. Thus, the process consists of a chain or network of interactions (events and responses). By bringing together a diverse group of managers, (potential) users and developers, a maximum output of events from the different perspectives is expected without a major bias of presumption that might exist if only people with the same role and perspective drove the modeling. Potentially, this method can be used for any kind of process modeling, but it has been developed for business process modeling and requirement engineering, e.g. in event-driven software design. In DiSSCo, the method was applied for quite specific work processes with CMS and the DiSSCO RI. Thus, the DiSSCo event storming workshop was organized for a mixed group of curators, collection managers, CMS users and vendors. For the representatives of the collections no deeper knowledge on technical topics was required. Likewise, the developers and vendors of CMSs and the DiSSCo RI were not expected to know all the details of the workflows within the collections. However, usually a high overlap in comprehension is the case as most of the participants work in a collection management context.

The overall goal of the Event Storming workshop as preparation for the DiSSCo pilot was to brainstorm and aggregate all kinds of events that could occur to a Digital Specimen in both the CMS and the DiSSCo RI.

After a session of several introductory talks on DiSSCo, openDS and the method used, the 33 participants were divided into three breakout groups. Each breakout group had a focal topic: "Development of a CMS or related system", "Usage of the DiSSCo Research Infrastructure" and "Usage of collection management systems". The allocation of the people to the groups was leveraged by the participant's individual interests and role within their institution indicated in the registration form of the workshop.

Facilitated by dedicated moderators (members of the WP 6 task group team) the participants were asked to use a virtual whiteboard on https://miro.com that contained a prepared structure and a summary of instructions for the collaborative work. They were asked to list the events they considered relevant, based on their related work routines. In addition, the collection of user stories describing evolving requirements of stakeholders involved in managing and using natural science collections. The user stories have been collected in the

ICEDIG project (https://github.com/DiSSCo/user-stories/issues) and have been expanded in DiSSCo Prepare (see deliverables D1.1 and D1.2 (Fitzgerald et al. 2021; von Mering et al. 2021)).

After the breakout sessions, the groups presented a summary of their results to the plenary. At the end of the workshop the participants were asked to assign stars (5 stars per person) in order to indicate their personal preference to the collected events. The individual results of the three breakout groups showed some overlap in the events listed, because people in different groups had similar ideas. In the post-processing of the workshop results, equal or very similar events have been aggregated to a unique list of events across the groups (see Appendix 1). The stars assigned to these events have been added to the aggregated events as well. In the final and aggregated outcome, the events with the most stars represent a ranking that can be considered as a priority list for the DiSSCo pilot.

3. Results

Results from the Event Storming Workshop

According to the aggregated results from the breakout groups, the most important events in a CMS relevant for the DiSSCo RI are related to specimen records (Fig. 1). The event *"Specimen received a new name"* received the highest priority, triggered by a researcher who "*checks literature in respect to nomenclatural code rules*" and/or "*classification updates*". The event "*New specimen record created in a CMS*" received the secondary priority, triggered by activities related to collection management e.g. "*Digitisation of existing historical specimen*", "*Specimen formally accessioned*" or "*Specimen gifted*" which are issued by a "*Collection manager*". The next-ranking event was *"Specimen re-determined in CMS"*, which is triggered by a taxonomic revision conducted potentially by different collection agents. In this case this event could also be triggered by a machine as an agent (e.g. by pattern recognition).

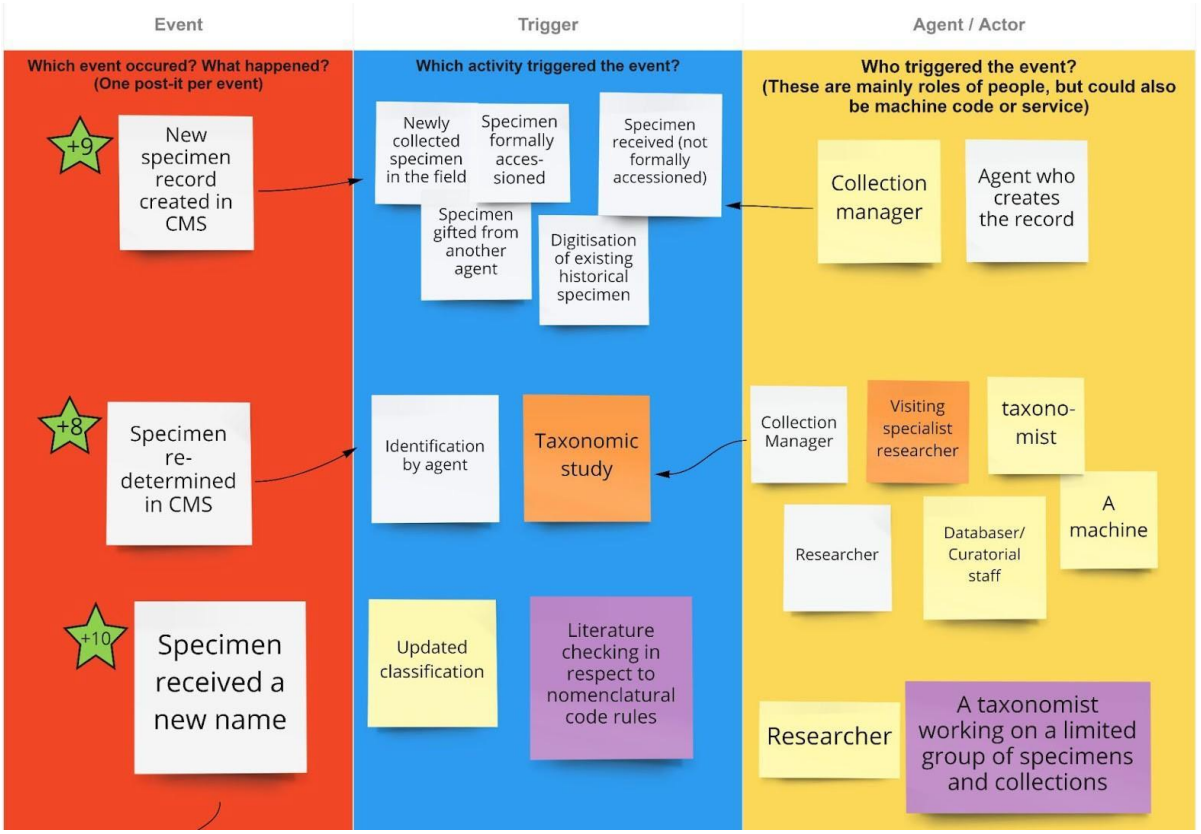


Figure 1 : Screenshot from the aggregated results (partial) from the workshop breack-out groups.

Appendix 1 holds a complete list of aggregated events identified in the event storming workshop. Independent from the priority identified, the events can be classified as members of event types, each of which consists of a subject and a predicate. Four different predictions have been considered relevant for event types in local CMSs: (1) addition (*created*), (2) modification (*changed*), (3) transaction (*processed*) and (4) deletion (*removed*). If these predicates are combined with a controlled list of subjects derived from events that could occur in a CMS, then generalization as a content abstraction layer can be accomplished to formally describe the event types independent from the original terminology within the different CMSs. This may indicate that an original (informal and CMS-specific) event description might need to be split up into different formal descriptions to meet the layer of abstraction. Table 1 contains a list of controlled terms that was derived from the event storming workshop as a basis for the pilot implementation and API recommendations.
For example, the top priority events would be classified and formally described as shown in Table 2. In row 3 the original event description "*Specimen received a new name*" can be characterized by emitting two formal events: "*taxon created*" and "*object name changed*" provided that the new taxon name was not yet in the taxonomic reference list of the respective CMS, so it needs to be created, and the CMS does not support multiple alternative determinations, so the name would be changed instead.

subject	description
object	typically the physical collection object or physical specimen or material sample
object availability	the object's availability status to indicate if it is physically available for loans, digitization, analyses or any other kind of physical handling
object identifier	locally unique identifier of an object, including alternative catalog numbers and inventory numbers.
object name	the name or title of the object
object record	any other (non-priority) digital metadata about the object
taxon reference	object metadata on the nomenclatural concept for the object's classification in order to assign a (scientific) name
geographic reference	object metadata on the geo-location where the object has been collected.
literature references	object metadata on the literature used or associated with
media record	media files and metadata associated with the object.
….	

Table 1: List of terms to be used as subjects for the formal description of events in CMSs.

original event description	formal description	
	subject	predicate
New specimen record created in a CMS	item	created
Specimen re-determined in CMS	identification	created
Specimen received a new name	taxon reference	created
	item name	changed
…		

Table 2: Example for mapping the events from a CMS to a formal description using controlled terms.

Document conventions

This document outlines requirements and recommendations for web APIs exposed by modules and services for integration in the DiSSCo RI. The following conventions are applied to distinguish between mandatory and optional features of DiSSCo-compliant web APIs in accordance to RFC 2119 (Bradner 1997):

- MUST - the usage of this term indicates features of the standard that any implementation is required to fulfill in order to be considered DiSSCo-compliant.
- SHOULD - indicates optional features that are highly recommended for implementation, but are not required; if these features are implemented they MUST follow the recommendations outlined in the standard.
- COULD - indicates optional features that are considered beneficial for the service, but are not required; if these features are implemented they MUST follow the recommendations outlined in the standard.

Recommendations

Considering the above-mentioned challenges the recommendations and guidelines are based on the following premises:

1. The connection between DiSSCo RI and CMS must be reliable and flexible;
2. To create a business model that digitally unifies all European natural science assets under common access and ensures that all the data is easily Findable, Accessible, Interoperable and Reusable (FAIR principles).
3. To enable a sustainable integration of external modules and systems for collection management into the DiSSCo RI.

Web-based CMSs are likely to implement machine-readable interfaces as REST APIs. Therefore an extension of (potentially existing) APIs with a few DiSSCo specific endpoints is a preferable solution for implementation. On the other hand, CMSs that are not web-based would have the option to set up a light-weight wrapper service to expose the DiSSCo specific API endpoints. As a second option, CMSs could directly implement the DOIP in order to be compliant with the Digital Object Architecture, but this is considered more difficult regarding the challenges of evolving systems (see above "*Challenge 6*"). Any time the DiSSCo RI is developed further, the CMSs would have to check the implicit compatibility with the latest developments. With a REST API designed for abstract events emitted, the only modifications in CMSs would be additional event types without any changes in the API endpoints. The support of event types would be limited to the CMSs features. Thus, the

DiSSCo API would only be modified content-wise as soon as new features are implemented in the CMS. Downward compatibility to certain API versions can be realized on the DiSSCo RI side of the connection instead of asking the CMS developers to change the code so regularly. Instead, a DiSSCo ingestion component will be maintained by the DiSSCo RI, to act as a broker between the CMS and the DiSSCo RI and allowing for pulling and pushing events and the respective payload using the DiSSCo API endpoints of the CMS. The ingestion component also takes care of the communication to the Digital Specimen index and registry in a DOA compliant manner (Fig. 2 and 3).
The overall aim for the harmonized integration of the CMSs into the DiSSCo RI should be for the developers of CMSs to have a minimum number of complex, technical tasks, so the focus can be on the useful content-oriented connection to DiSSCo.

API Guidelines

These API guidelines are recommended to be implemented into a CMS in order to allow for its connection to and integration into the DiSSCo RI.
The DiSSCo REST API guidelines adhere to the JSON API - specification version (v1.0).

Endpoints expected in CMSs for DiSSCo integration

Table 3 lists DiSSCo endpoints for a JSON:API compliant API implementation in a CMS.

Endpoint	Description
GET /dissco/	returns OpenAPI-compliant document for machine-readable API documentation
GET /dissco/doc	(optional) returns a human-readable representation of the API documentation.
GET /dissco/capabilities	returns metadata on the dissco endpoint (e.g. name of the CMS, list of event types supported by the respective CMS)
POST /dissco/auth	endpoint to request an authentication token (e.g. OAuth 2.0)
GET /dissco/event/{id}	endpoint for reading events and event metadata from the CMS structured according to CloudEvents specification. If no id is provided, a list of paginated events is returned
POST /dissco/event	endpoint for sending events to the CMS

Table 3: Suggested endpoints for the implementation of API endpoints in a CMS.

It is assumed that the DiSSCo-specific endpoints must not be in conflict with endpoints that may already exist in the CMSs native API (if applicable). Thus, the path "***/dissco/***" was chosen to distinguish it from other resources. Calling the empty endpoint should return the OpenAPI-compliant specification of the API implemented, which gives the opportunity to have a machine-readable documentation as close as possible. This could be complemented by a human-readable documentation available at "***/dissco/doc/***". Using existing tools that convert the OpenAPI YAML or JSON file into an HTML page (like e.g. Swagger UI) is the easiest way to achieve this and to keep the machine- and human-readable documentations in sync.

The endpoint "***/dissco/capabilities/***" must return the most important CMS-specific metadata (e.g. name, version) and - most importantly - the list of supported event types as this may vary in each CMS.
For example:


```
{
  "links": {
    "self": "https://collection.myinstitution.com/dissco/capabilities",
  },
  "data":
   {
        "type": "capabilities4dissco",
        "id": "https://collection.myinstitution.com/dissco/capabilities",
        "attributes":
        {
             "cms": {
                  "name": "My Collection Management System",
                  "version": "10.1.7",
                  "contact": "someone@example.com",
             },
             "supportedeventtypes": [
                  "org.dissco.event.object.created",
                  "org.dissco.event.object.changed",
                  "org.dissco.event.object_name.created",
                  "org.dissco.event.object_name.changed",
                  "org.dissco.event.object_name.removed",
                  "org.dissco.event.object_record.changed",
                  "org.dissco.event.media_record.created",
                  "org.dissco.event.media_record.changed",
                  "org.dissco.event.media_record.removed",
```

```
            ]
        }
    }
}
```

The endpoint "***/dissco/event/***" must return a JSON:API-compliant response which allows paginated browsing of the events as long a no specific event `id` is added to the endpoint ("***/dissco/event/{id}***") for directly calling the resource of a particular event (e.g. "*/dissco/event/A234-1234*"), the `type` of data in the JSON:API response must be "`event`" and the property `attributes` must be a JSON object compliant to the CloudEvents specification.

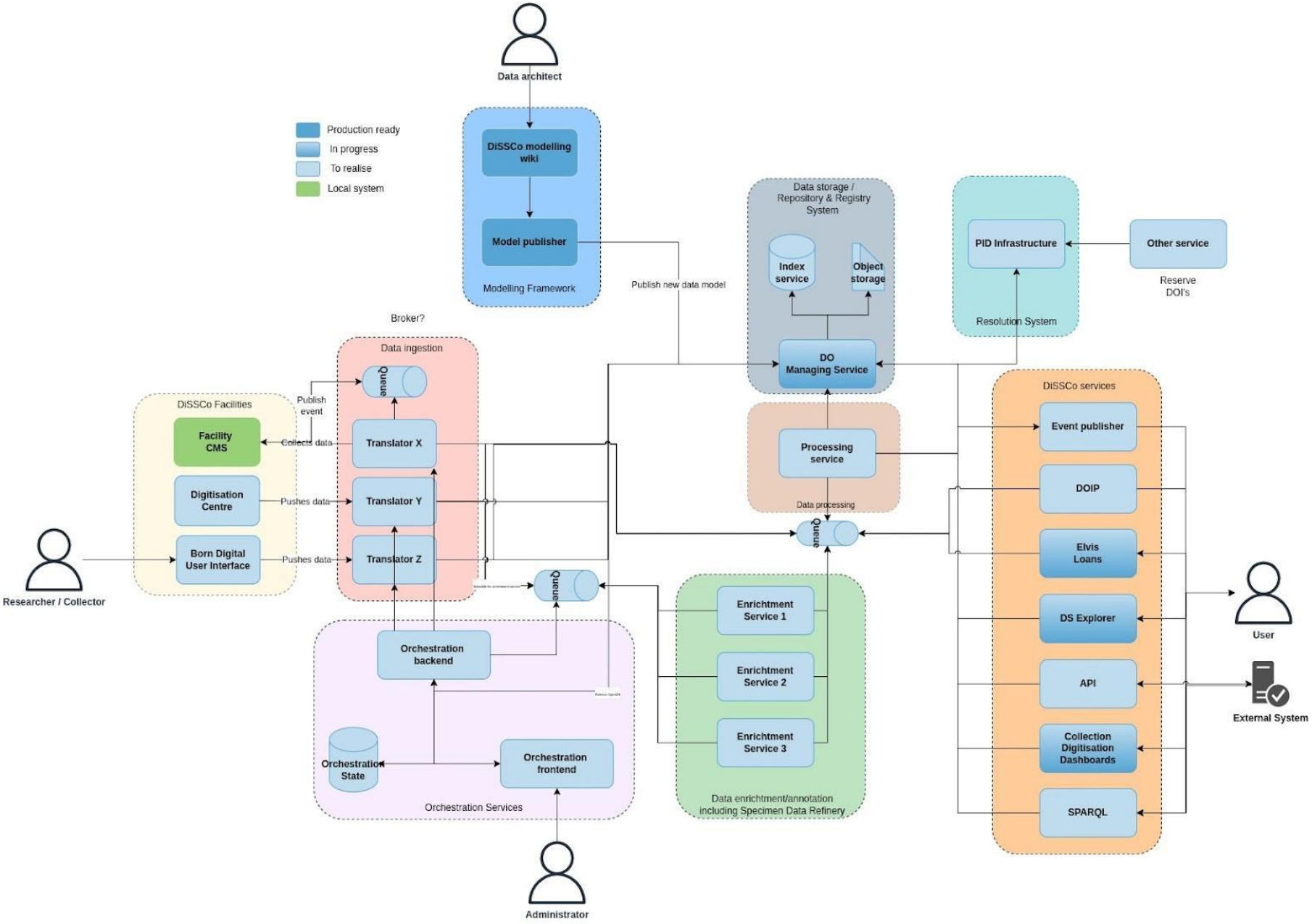
Example of a CloudEvent resource nested in a JSON:API-compliant resource:

```
{
  "links": {
    "self": "https://collection.myinstitution.com/dissco/event",
    "next":
    "https://collection.myinstitution.com/dissco/event?page[offset]=2",
    "last":
    "https://collection.myinstitution.com/dissco/event?page[offset]=10"
  },

  "data":
  [
       "type": "event",
       "id": "A234-1234",
       "attributes":
       {
            "specversion": "1.0",
            "type": "org.dissco.event.object.created",
            "source": "https://collection.myinstitution.com/dissco/event",
            "subject": "item 123",
            "id": "A234-1234",
            "time": "2022-02-06T17:31:00Z",
            "datacontenttype": "text/json",
            "data": "{ // put the payload here }"
       },
       "links": {
         "self": "https://collection.myinstitution.com/dissco/event/A234-1234"
       }
  },
  {
       "type": "event",
       "id": "A234-1234-1235",
```

```
        "attributes":
        {
             "specversion": "1.0",
             "type": "org.dissco.event.object.changed",
             "source": "https://collection.myinstitution.com/dissco/event",
             "subject": "item 123",
             "id": "A234-1235",
             "time": "2022-02-06T18:35:00Z",
             "datacontenttype": "text/json",
             "data": "{ // put the payload here }"
        },
        "links": {
          "self": "https://collection.myinstitution.com/dissco/event/A234-1235"
        }
    }
}
```

Figure 2 : General DiSSCo Architecture overview, including the CMSs and DiSSCo RI published events.



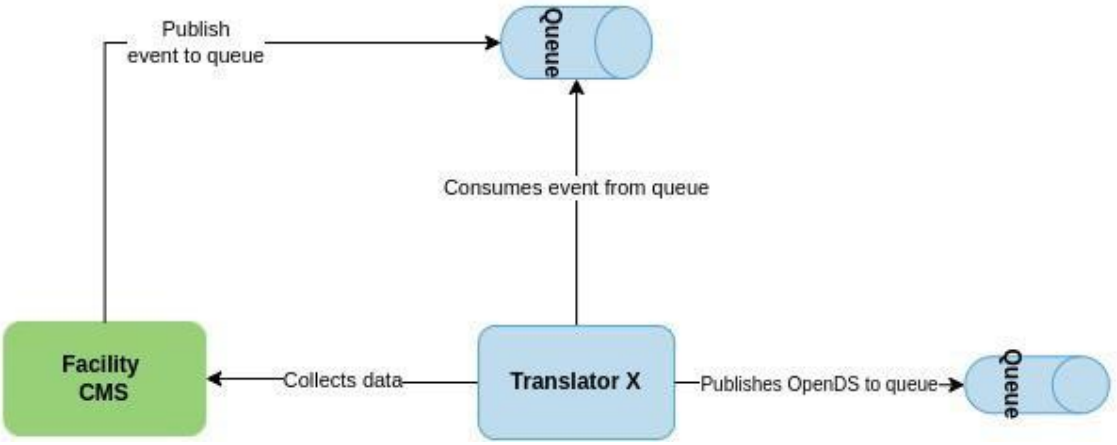


Figure 3 : Detailed view on the ingestion event queue of the General DiSSCo Architecture.

4. Discussion and Outlook

Due to the high diversity of CMSs, an integration into DiSSCo can only be successful with a harmonized abstraction layer. In order to achieve this abstraction existing standards and interfaces could be used. However, unresolved challenges might remain.
Using data publication tools such as the BioCASe Provider Software (BPS) and the Integrated Publishing Toolkit (IPT), the one-directional way of specimen data could be handled quite well. These tools are widely distributed in the community of natural history collections for providing data to portals like GBIF. The used data standards (Darwin Core and ABCD) are designed for the exchange of occurrence and collection data. However, IPT and BPS in the current state would not overcome the challenge of synchronization (see Challenge 1) between the records living in local CMSs and the DiSSCo RI, because they are designed to only provide data and thus, lack endpoints for receiving data. They access the databases of the CMSs but they are not integrated in the CMS workflows besides the final step of publishing a snapshot of collection data. Thus, data receiving and/or actively pushing data to the diverse CMSs would need an essential amount of work in the code bases of both the data provision tools and additionally the CMSs, as the latter would at least need to allow incoming, external data to be processed and stored.

In order to serve an abstraction layer in a useful manner, a direct implementation of interfaces appears to be the most promising approach to reach technical readiness of CMSs for connection and integration into the DiSSCo RI. This could be done by implementing the DOIP in the CMSs to be compliant with the DiSSCo's Digital Object Architecture. However, DOIP introduces a new protocol and complexity in CMSs for functionality that could be realized in an easier way in order to better address the challenges of existing (developer) ressources (see Challenge 5) and the technological changes (see Challenge 6). To date, the software component Cordra (version 2.4) used for the NSIDR does not fully implement the DOIP and uses internal REST API calls instead. Thus, implementing DOIP in CMSs would currently not leverage technical readiness for DiSSCo. Instead, using and developing REST APIs is common practice. REST APIs are well-known and can be considered established state of the art. Thus, developers of CMSs would not need to gain any special knowledge on a new protocol. Furthermore, some web-based CMSs do already have a REST API, so an extension of these would presumably be the easiest way to implement DiSSCo-specific functionality.
Nevertheless, there are CMSs without REST APIs and without any web technology used. These would need to be enabled for web-communication with other systems in any case to

be ready for DiSSCo connectivity. Therefore, institutions with offline CMSs might be candidates for one of the three options: 1) using a web-based wrapper tool that includes the DiSSCo endpoints, 2) a change of CMS or major reconstruction of the CMS used, or 3) using a centralized CMS as a service connected to DiSSCo.

In order to overcome the challenge of diversity of CMS REST APIs, the only efficient approach is using existing, agreed standards (such as JSON:API and OpenAPI) together with a set of DiSSCo-specific design guidelines. The special requirement for this is that the DiSSCo endpoints must not interfere with existing API endpoints in the CMS and they need to be generically usable, so the name and version of the underlying CMS must not be relevant for the ability to communicate with the DiSSCo endpoints. This can be achieved with the event-driven approach (Islam 2019). Following the assumption that everything that happens in the CMS and DiSSCo RI can be described as an event, a stable content-wise abstraction layer can be generated with fixed API endpoints. Even if there are some new features indicating new or different behavior such as responses or reactions to a specific event, the API doesn't have to be changed at all. Only new event types need to be introduced to be highly flexible. However, development might be needed in the local CMS in order to implement specific response procedures triggered by new event types. But this still allows for high flexibility in terms of resources (Challenge 4) and prioritization (Challenge 5), because new event types only need to be addressed if they are considered relevant for the workflows, use cases and needs in a particular CMS. In general, a certain event type would be understood and its respective payload would be processed by a CMS or not without breaking the API compatibility. In comparison to the possible pressure to always update a piece of software to the latest version in order to make use of its latest features, new event types would not break the connection between the CMS and DiSSCo even if an older version was used. Thus, there is no strong dependency as long as no fundamental features and mechanisms such as the API authentication methods, aspects of cyber security or the specification of the underlying event schema are changed.
It is of special importance that the definition of event types is being pushed forward and consolidated. The Event Storming workshop could only initiate the process of identifying event types for the pilots. To fully address the priorities of the community (see Challenge 5) more user stories from different stakeholders should be included. However, for the prototypes of implementing DiSSCo API endpoints in a CMS the priorities as outcome of the workshop are very useful. Likewise, the payloads need more detailed specification as the definition of Open Digital Specimen Specification (openDS) proceeds. Which means that the CMSs not only need to understand the event types on the abstraction layer, but also they need to be capable of processing the structure of the events' payload. The payload could

be one or many identifiers and metadata to be used to fetch the actual data in a subsequent step. This would be advantageous especially for large data packages as the events would be rather small messages without a huge overhead of data. On the other hand, the payload could also be the data itself structured in an (event type specific) well-known format, such as a serialized JSON for openDS which makes use of ABCD, DarwinCore, DublinCore, MIDS and other ontologies. This would introduce the advantage of keeping the event related data very close to the event's metadata. As soon as a sequence of micro-changes triggers a lot of events in a short period of time (in terms of only a few (milli-)seconds), this approach would avoid additional (CMS specific) API calls.

This deliverable describes a concept for the integration of collection management systems into the DiSSCo Research Infrastructure, including recommendations for APIs guidelines. But it does not provide final blueprints for the construction of the DiSSCo API endpoints for CMSs. The upcoming pilots that make use of the results of this deliverable will finally highlight the weaknesses and strengths of the suggested approach. Considering this deliverable as a first version of the harmonization concept for CMSs with DiSSCo, the experiences of prototypic implementation will be fed back in the subsequent version in order to approve or revise the above-mentioned guidelines and recommendations. In the meantime, the controlled vocabulary and list of event types will be finalized and consolidated in order to leverage the formal description of events relevant for CMSs and the DiSSCo RI.

Once the prototype practically demonstrates the suggested approach and some possible improvements for the DiSSCo REST API could be derived from it, the first release of the API specification should be published. As soon as several CMSs implement the DiSSCo API guidelines, a technical team needs to keep track of the compatibility while both the CMSs and DiSSCo RI evolve. For this a certification procedure with standardized compliance tests need to be established, so CMS vendors could tag their CMS as “DiSSCo ready”.

5. References

Access to Biological Collection Data (ABCD). Available at: https://abcd.tdwg.org/

Addink W., Hardisty A. (2020) 'openDS' – Progress on the New Standard for Digital Specimens. Biodiversity Information Science and Standards 4: e59338. https://doi.org/10.3897/biss.4.59338

Bradner, S. (1997)"Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, https://www.rfc-editor.org/info/rfc2119 .

Brandolini, A. (2013-11-18). "Introducing Event Storming". https://www.eventstorming.com/book/. Retrieved 2017-04-06

CloudEvents Specification version 1.0.2. Available at https://github.com/cloudevents/spec/tree/v1.0.2, accessed 2022-02-16.

Corporation for National Research Initiatives (CNRI). Cordra Digital Object Repository and Registry software. Available at: https://www.cordra.org/.

Darwin Core. Available at: https://dwc.tdwg.org/

Dillen M., Groom Q. & Hardisty A. (2019). Interoperability of Collection Management Systems. Zenodo. https://doi.org/10.5281/zenodo.3361598

DONA Foundation. (2019). Digital Object Architecture. Available at: https://www.dona.net/digitalobjectarchitecture.

DONA Foundation. (2018). Digital Object Interface Protocol Specification. Version 2.0. https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf

FAIR Principles: https://www.go-fair.org/fair-principles/

Fitzgerald H., Juslén A., von Mering S., Petersen M., Raes N., Islam S., Berger F., von Bonsdorff T., Figueira R., Haston E., Häffner E., Livermore L., Runnel V., De Smedt S., Vincent S., Weiland, C. (2021). DiSSCo Prepare Project Deliverable D1.1 Report on life sciences use cases and user stories. https://doi.org/10.34960/xhxw-cb79

Glöckler F., Macklin J., Shorthouse D., Bölling C., Bilkhu S., Gendreau C. (2020) DINA—Development of open source and open services for natural history collections & research. Biodiversity Information Science and Standards 4: e59070. https://doi.org/10.3897/biss.4.59070

Güntsch A., Hyam R., Hagedorn G., Chagnoux S., Röpert D., Casino A., Droege G., Glöckler F., Gödderz K., Groom Q., Hoffmann J., Holleman A., Kempa M., Koivula H., Marhold K., Nicolson N., Vincent S., Triebel D. (2017). Actionable, long-term stable and semantic web compatible identifiers for access to

biological collection objects. Database : the Journal of Biological Databases and Curation. 2017(1). https://doi.org/10.1093/database/bax003.

Hardisty A. (2018). Widening access to European natural science collections with Digital Specimens and Natural Science Identifiers (NSId). https://alexhardisty.wordpress.com/2018/11/29/widening-access-to-european-natural-science-collections-with-digitalspecimens-and-natural-science-identifiers-nsid/. Accessed on: 2020-9-11.

Hardisty A. (2020). What is a Digital Specimen? https://dissco.tech/2020/03/31/what-is-adigital-specimen/. Accessed on: 2020-9-11.

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(1), p.50. DOI: http://doi.org/10.5334/dsj-2020-050

Islam, S. (2019). Event-driven digital object architecture for natural science collection. https://sharif-islam.medium.com/event-driven-digital-object-architecture-for-natural-science-collection-160a3b1f3044 Accessed on: 2022-02-25.

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

Lendemer J., Thiers B., K Monfils A., Zaspel J., R Ellwood E., Bentley A., LeVan K., Bates J., Jennings D., Contreras D., Lagomarsino L., Mabee P., Ford L., Guralnick R., Gropp R., Revelez M., Neil C., Seltmann K., Aime M C., (2020). The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education, BioScience, Volume 70, Issue 1, Pages 23–30, https://doi.org/10.1093/biosci/biz140

Lannom L., Koureas D., Hardisty A., (2020). FAIR Data and Services in Biodiversity Science and Geoscience. Data Intelligence. https://doi.org/10.1162/dint_a_00034

Petersen M., von Mering S., Pim Reis J., Glöckler F., Weiland C., Addink W., Cubey R., Güntsch A., Fichtmüller D., Dillen M. DiSSCo Prepare Project - Milestone Report 5.2 “Implementation of concepts for sustainability of services, CMS, and overall TRL”

von Mering S., Petersen M., Fitzgerald H., Juslén, A., Raes N., Islam S., Berger F., von Bonsdorff T., Figueira R., Haston E., Häffner E., Livermore L., Runnel V., De Smedt S., Vincent S., Weiland C. (2021). DiSSCo Prepare Project Deliverable D1.2. Report on Earth sciences use cases and user stories. https://doi.org/10.34960/n3dk-ds60

Appendix 1

List of aggregated events identified in the Event Storming workshop (ordered by ranking / user priority).

Event	Trigger	Agent	Number of stars
Specimen received a new name in CMS	Taxonomic study	Collection Manager	10
New specimen record created in CMS	Newly collected specimen in the field	Collection Manager / Curatorial staff	9
Specimen re/determined in CMS	Taxonomy study	Collection Manager / A machine	8
Creation of names not linked to published literature	Mass digitization project	Database stuff / curatorial stuff	6
Augmentation of digital record from secondary source	Literature research	Curator	6
Term remapped to new / additional vocabulary	new mappings & crosswalks available	Community	5
Georeference corrected	Algorithm correcting georeferences	Bot	5
Specimen imaged	Request for image / Image created	Researcher / Collection manager	5
Physical object moved to other collection	Transfer of collection to another institution	Political decisions	4
An annotation added	Identified an error in the data	Citizen scientist	4
Data digitized	Specimen data refinery job request	Head of collections	4
Publishing a specimen	Publication	Researchers / students / curators	3
Augmentation of digital record from primary source	Schedule inventory process	Curator / collection manager	3
Subdivision of object lot into individuals	Re-curation or re-identification of individuals	Curator	3

Specimen record merged	Duplicate records discovered	“Machine” automatic consistency check	2
Updating index	Indexing metadata for statistics	A Service	2
New identification added	Making an identification	Taxonomist / machine	2
Georeference added	Georeferencing project	GIS digitiser	2
Link collectors to Binomia / Wikidata	Data improvement	Data managers / curators	2
Personal data removed from specimen	Requested personal data to be removed	Agents who created the record	1
Specimen record reverted to earlier version	User noticed error / reverted a change	Collection manager	1
Specimen georeferenced in CMS	Locality georeferenced	Human / Machine (GeoLocate)	1
Information added to existing record	Digitization from sources materials	Collection manager	1
Specimen analysed	Analytical data was gathered	Researcher / Collection manager	1
Label data transcribed	Request for data	Researcher	1
New genetic sequence of specimen added	New sequence lodged with Genbank or BOLD	Researcher / Collection manager	1
Specimen was loaned to another institution	Loan request	Researcher / Exhibitor of the Museum	1
Specimen was sampled	Specimen sampled requested	Researcher	1
Specimens were exchanged between two institutions	Collaboration / exchange request	Researcher	1
Specimen measured	Research event	Researcher / machine	1
Loan request	Taxonomic study	Visiting specialist researcher	1
Researchers asked a question about a facility	Question submission through the portal	Researcher	1

A list of specimen returned	a search for specimen meeting criteria	Researcher	1
An old, ambiguous locality name was changed from “A” to “B” since the correct origin of series of specimens was proved to belong to a different country	Initiative to resolve ambiguity among similar sounding locality names	The National Geographic Society of a certain Country	1
Object loaned	Loan request received	Researcher	0
New citation of specimen added	Machine PID identification	Machine citation ID sync	0
Classification changed due to new information	New information through publication	Researcher	0
Specimen donation was received and recorded to a database	Specimen donation	Anyone (private collection owner, another institution)	0
Specimen gifted to another agent	Distribution of duplicates	Collection manager	0
Specimen condition assessed	Digitisation / Curation event	Collection manager	0
Specimen conserved	Condition assessment	Digitiser / Curator	0
Seeds were sent to another botanic garden and recorded as a material transaction	Seed order	Living collections curators	0
Tissue sub-sample taken	Tissue taken from specimen	Field researcher/ Collection manager	0
Accession new collection	Collection arrived on site	Collection curator	0
Derivative object created (e.g. tissue sample / thin section)	Research request	Internal / External scientist	0
Taxonomic reclassification	Publication	Researcher	0

Associate an object to related object (e. g. galls on branch, multiple fossil taxa on slab, distributed collections)	Research activity induces need to document relationships	Internal / External scientist	0
Specimen conservation treatment	Laboratory procedure	Conservator	0
Specimen cited in literature	Publication of new monograph, taxonomic work	Researcher	0
Link multiple sheets of the same specimen	Data cleaning	Data managers / curator	0