H2020-INFRADEV-2019-2
Grant Agreement No 871043

DiSSCo Prepare WP6 D6.4:

Implementation of the DiSSCo Data Management Plan (DMP) and ENVRI FAIR compliance of DiSSCo data services.

Work package lead: Claus Weiland

Authors:

Claus Weiland	0000-0003-0351-6523	Senckenberg Nature Research Society
Wouter Addink	0000-0002-3090-1761	Naturalis Biodiversity Center
Mathias Dillen	0000-0002-3973-1252	Meise Botanic Garden
David Fichtmueller	0000-0002-0829-5849	Botanic Garden and Botanical Museum Berlin
Jonas Grieb	0000-0002-8876-1722	Senckenberg Nature Research Society
Elspeth Haston	0000-0001-9144-2848	Royal Botanic Garden Edinburgh
Mikko Heikkinen	0000-0001-9898-1916	Finnish Museum of

		Natural History Luomus
Sharif Islam	0000-0001-8050-0299	Naturalis Biodiversity Center
Sam Leeflang	0000-0002-5669-2769	Naturalis Biodiversity Center
Esko Piirainen		Finnish Museum of Natural History Luomus
Julia Pim Reis	0000-0003-4383-0357	Museum für Naturkunde Berlin
Tim Robertson	0000-0001-6215-3617	Global Biodiversity Information Facility
Hanieh Saeedi	0000-0002-4845-0241	Senckenberg Nature Research Society

Version:

Version	Date	Contributors	Comment
0.1	2022-10-15	Claus Weiland	Draft Iteration of the *Provisional Data Management Plan for DiSSCo infrastructure* w.r.t. maDMP
0.2	2022-12-04	Sam Leeflang, David Fichtmüller, Jonas Grieb, Sharif Islam, Mikko Heikkinen, CW	Iteration step detailing implementation of the maDMP (PROV-O/RFC6902) and (DCAT) mappings
1.0	2022-12-12	All authors	Finalizing for submission

Table of Contents

Preface

The foundations for DiSSCo's data management plan were developed under the leadership of Alex Hardisty in the framework of the "ICEDIG - Innovation and consolidation for large scale digitisation of natural heritage" project. The *Provisional Data Management Plan for DiSSCo infrastructure* was conceived as a living document (Hardisty 2019), i.e. a document that is to be continually supplemented and updated based on the technical advancement of DiSSCo's data models, services and architecture as well as adjusted in response to developments in the global and in particular Europe's scientific and technological landscape rather than completely revised.

The amendments presented in this report essentially result from two related developments: The increasing elaboration of DiSSCo's core data model, the Digital Extended Specimen (Islam 2020, Hardisty 2022) and the further clarification and specification of the FAIR principles with regard to the implementation of FAIRs key concept "machine actionability" (Jacobsen 2020, Lannom 2021) in the context of data management plans (Miksa 2019).

Keywords

Distributed System of Scientific Collections, DiSSCo RI, Machine-Actionable Data Management Plan, Research Data Management, FAIR, Machine-Actionability, Digital Extended Specimen, FAIR Digital Object, Data Catalog Vocabulary. PROV-O

Abstract

Data Management Plans (DMPs) are documents guiding actions and liabilities of research data management during the whole data lifecycle, including project outputs and involving the maintenance of correct attribution and provenance data. Following the approach of FAIR Digital Objects (FDOs) adopted by DiSSCo, we developed a provenance data model which preserves the operations acting upon DiSSCo’s core data model, the Digital Extended Specimens (DES). A DES is a specific FDO version which represents a particular geo- and biodiversity specimen in a natural science collection and binds - i.e. provides encapsulated links or bitstreams of - relevant derived information content about that physical specimen.

Based upon the key objective of the FAIR principles referred to as “machine-actionability” - the capability of software agents or “machines” to handle data autonomously and appropriately with little or without human intervention - we adapted the concept of a “machine-actionable data management plan” (maDMP). The maDMP for DiSSCo is designed as a continuously self-updating structure which is an FDO itself and offers significantly more options over the traditional approach of research data management (RDM) driven by static documents and reports. The maDMP enables flexible integration of all relevant information in the data life cycle as well as the creation of downscaled reports containing structured data. Consequently, it facilitates data discovery and reuse, and supports automated evaluation and monitoring.

Within this report, we present a prototype as part of DiSSCo’s specification for open Digital Specimen (openDS) which builds upon the openDS classes for provenance recording. This is supplemented by additional concepts from the W3C PROV Ontology (PROV-O) and type-specific operations providing timestamped records using JSON Patch (RFC 6902).

To foster integration and reuse of research data managed in DiSSCo, we present an initial approach for the semantic alignment of the provenance and maDMP classes with corresponding terms in the W3C Data Catalog Vocabulary.

On this basis, we discuss the relationship of approaches providing an overarching ontology as mapping target - one common ontology maps to many provider vocabularies - vs. the FDO approach using a common set of globally obtainable (Kernel) attributes - many-to-many mappings of different provider/registry standards using an agreed upon framework of specified types, profiles and Kernel Attributes - and compare both concepts with regard to achieving interoperability in hyper infrastructures like the European Open Science Cloud.

Introduction

The European Research Initiative Distributed System of Scientific Collections (DiSSCo) is a dynamic network of research infrastructures from the area of natural science collections (NCS). Today (December 2022) there are more than 170 DiSSCo partner institutions located in 23 countries.

Key objectives for DiSSCo's distributed technical architecture are improved capabilities for the discovery of European NSC's data, harmonized and facilitated digital access to these collection data, and the provision of a service-layer for enhanced interpretation, curation, annotation reuse and repurposing of specimen data both by humans (in form of community-based curation) and increasingly by software agents or in short “machines” (Mons 2020). These objectives reflect the transformation of NSCs from curio cabinets into vast repositories of biodiversity and geological specimens acting as hubs for the corresponding geo- and biodiversity knowledge (Meineke 2018). Major challenges for this transformation process are the development of interoperable data standards and models as well as building up the information infrastructures for data exchange between the different research fields and newly established regional clusters of biodiversity data providers including USA's Integrated Digitized Biocollections (iDigBio), the National Specimen Information Infrastructure (NSII) of China and Australia's digitisation of national research collections (NRCA Digital) (Hardisty 2019b). This initiated a harmonization and convergence process in the biodiversity community resulting in the Digital Extended Specimen (DES) concept (Webster 2021).

A DES in the scope of DiSSCo encapsulates and links persistently to all relevant information artifacts, which are about the corresponding physical specimen in a collection (Figure 1).

Thus a representation of the logical structure of a specimen's data is implemented, enabling operations like annotations of the entire object though it consists of distributed components. Several specifications are of particular relevance for DES: The Minimum Information about a Digital Specimen (MIDS) specifies and classifies the relevant information elements assignable to a specimen within a digital framework (MIDS 2022). An assigned MIDS level from zero to three informs about gaps and completeness of the description.

The open Digital Specimen (openDS) specification outlines how to implement the DES data model by binding all necessary information (metadata like rights, permissions, licenses as well as links to content) about an entity into a composite digital object. Both the embedded components in a complex digital object (e.g. associated media types like 2D images) and the properties and relations directly used from them (e.g. the MIDS level) are described as openDS elements.

A further aim of the DES concept is to foster the reuse of specimen data in related disciplines like Life Sciences, Agrosystems Research or Earth System Sciences as well as in larger data and service federations like the European Open Science Cloud (EOSC) or the community of Environmental Research Infrastructures (ENVRI, Papale 2020, de Natale 2021).

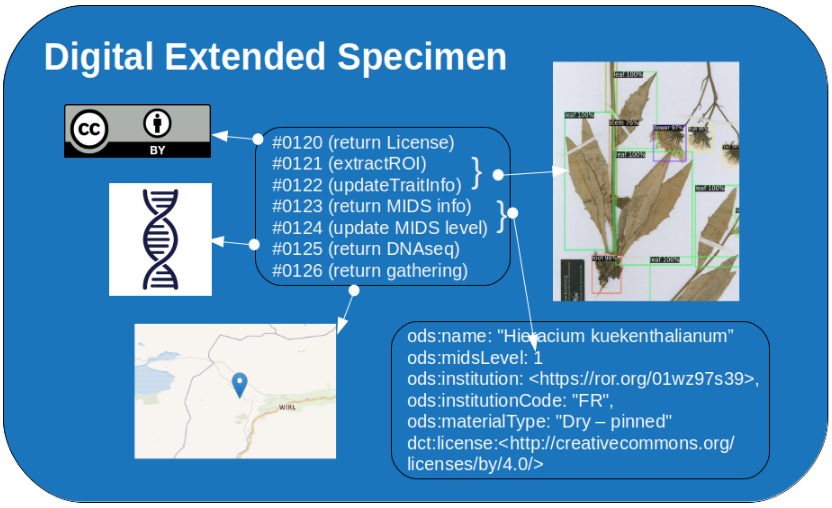

Digital Extended Specimen
BY
#0120 (return License)
#0121 (extractROI)
#0122 (updateTraitInfo)
#0123 (return MIDS info)
#0124 (update MIDS level)
#0125 (return DNAseq)
#0126 (return gathering)
ods:name: "Hieracium kuekenthalianum"
ods:midsLevel: 1
ods:institution: <https://ror.org/01wz97s39>,
ods:institutionCode: "FR",
ods:materialType: "Dry – pinned"
dct:license:<http://creativecommons.org/licenses/by/4.0/>

Figure 1: A Digital Extended Specimen as container object for related and derived information about a corresponding physical specimen in an NSC. The DES encapsulates descriptive metadata used for indexing, discovery and identification like geolocation data, technical and administrative metadata for rights and usage like licenses, primary data like images or sequences as well as typed information signaling which operations are appropriate for a DES. In the example shown here a DES containing herbarium scans is processed in a machine learning pipeline for the extraction of (morphological) trait features based on detection of regions of interest (ROIs) in a 2D image object (Younis 2020, Grieb 2021).

Namely the FAIR principles play a crucial role for the cross-domain reuse of research data by provision of clear definitions and metrics for good data management and stewardship for digital research assets (Wilkinson 2018). However, FAIRness of data cannot be achieved by information modeling of the data alone, but has to integrate an infrastructure perspective to comprise relevant services to reach FAIR-compliance (Collins 2018).

A key objective of the FAIR is to realize self-contained operation of machines on digital resources with little or without human intervention (Wilkinson 2016). The capability of machines to handle data autonomously and appropriately is a core objective of the FAIR principles which is often referred to by the term “machine-actionability” (Jacobsen 2020, Lannom 2021).

A prerequisite to achieve such autonomous data processing with minimized human oversight is the integration of the data into a network of FAIR-enabling services, which implement the required policies, rules, procedures and infrastructures (Schwardmann 2020).

An approach from the context of the Research Data Alliance (RDA) to realize a data model and a related infrastructure design for a suitable ecosystem of FAIR services are FAIR Digital Objects (FDOs) and the related Digital Object Architecture (DOA; Kahn 2006,

Wittenburg 2021). FDOs are embedded in a FAIR ecosystem of services for creation and maintenance of FDOs including data type registries, services to mint PIDs, and terminology services to provide FAIR-compliant cross-references to semantic artifacts like vocabularies and ontologies (Collins 2018).
This is accomplished through a set of specifications recapped in a unified conceptual model, comprising the essential components of any FDO: (i) The globally unique persistent identifier (PID) resolvable to essential metadata (called Kernel Attributes), (ii) the model layer providing a profile defining how machines can autonomously operate on the FDO and (iii) the resource layer providing actual content including typed data structures, contexts, and policies of the data producers (Figure 2,).

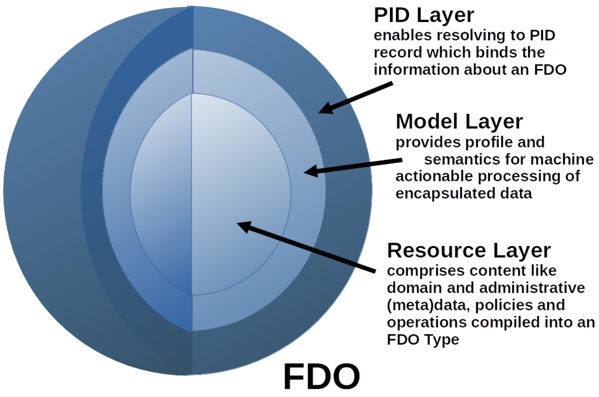

PID Layer
enables resolving to PID record which binds the information about an FDO
Model Layer
provides profile and semantics for machine actionable processing of encapsulated data
Resource Layer
comprises content like domain and administrative (meta)data, policies and operations compiled into an FDO Type
FDO

Figure 2: Unified conceptual model showing the essential structural parts of a FDO: (i) PID Layer, providing a globally unique resolvable and persistent identifier (PID) resolvable to PID a set of so-called kernel attributes which enable persistent, secure and trusted access to the FDO, (ii) the FDO model layer that defines the semantics for machine actionable reuse and access of that object in form of attribute sets (called FDO Profiles) and (iii) the resource layer that encapsulates the actual content of the FDO in form of enclosed bitstream and linked resources in a distributed landscape of trusted and FAIR compliant service providers.

Based on the flexibility and conciseness of this conceptual model, an appreciable number of domain-specific implementations was achieved (Wittenburg 2022). It should be noted that PID and model layer are to a large extent specified by the major FDO-community organization, the FDO Forum. Exemplarily, (Hardisty 2021) describes DISSCo's selection process of a persistent identifier scheme for the DES within the framework of these specifications.
Substantiating the FDO approach for a particular domain therefore means to characterize signals (signs) indicating potential self-contained operations of machines on the data allocated in the FDO's resource layer - such a set of characteristic signs forms the so-called FDO Type.
Accordingly, the major role of openDS is to provide a formal specification (i.e. an ontology; Gruber 2009) of the categories/classes, properties and relationships used to model the entities in context of NSCs, like specimen and media objects, in compliance with the FDO specifications with particular regard to the (yet not publicly released) methods for defining and attaching types to FDOs.

Data Management Plans and Provenance

Data Management Plans (DMPs) are documents guiding actions and liabilities of research data during the whole data lifecycle including project outputs and involving the maintenance of correct attribution and provenance data (Miksa 2019).

For the composition of DMPs, the “Ten Simple Rules for Creating a Good Data Management Plan” are often referred for the provision of good guidance for comprehensive DMPs (Michener 2015). Starting from these principles, a transformation from the understanding of the DMP essential as a kind of static document towards a more dynamic digital object, which integrates all relevant information in the data life cycle was proposed (Miksa 2019). In this concept, the DMP is now machine-actionable, which means it receives updated input from tools and services involved in the data processing pipelines.

Based on the results from preceding projects like ICEDIG, joint discussions in the openDS breakout group and community events like the DiSSCo Prepare Round Tables or the Consultation on the Convergence of Digital and Extended Specimens (Alliance for Biodiversity Knowledge 2021), the following essential characteristics for the data management of Digital Extended Specimen were prioritized (Leeflang 2022):

- DES is the core component and the primary digital object type of the DiSSCo architecture
- Accuracy and authenticity of the DES
- FAIRness
- Protection of data (legal regulations and community norms)
- Preserving readability and retrievability
- Traceability (provenance) of specimens
- Annotation history
- Determinability (status and trends) of digitisation
- Securability (authentication, authorization, accounting, auditing)

Focusing on the core data objects of DiSSCo, this requires as a condition that rich provenance data will be generated and preserved by all operations acting upon a DES, including timestamped records of change to enable reconstruction of a specific state at a date and time in the past (Hardisty 2019).

In common with other authors (Khan 2019), we understand provenance as defined by the World Wide Web Consortium (W3C):

“Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability, or trustworthiness." (Moreau 2013)

Following up on Khan et al. (Khan 2019), we can distinguish between *retrospective* and *prospective provenance*. Retrospective provenance retains all the records of operations and activities to achieve a specific state or result of a process together with rich contextual information. Prospective provenance in contrast describes the required sequence of operations to achieve a certain state or result in the form of a recipe of the computational steps, which can be applied to the same target to restore an earlier version, but also to similar data structures to update these (Davidson 2008).
To meet the aforementioned objectives, we associate both prospective and retrospective provenance information with a DES: Retrospective provenance guarantees the preservation of the annotation history and the determinability (status and trends) of a specimen’s digitisation while prospective provenance enables the traceability of specimens in terms of restartability and retrievability of earlier versions.

As described below, the PROV data model (PROV-DM; Belhajjame 2013) is used to capture retrospective provenance information. The elements for the event-based capturing of provenance information in the openDS framework are sketched in Figure 3 and will be detailed in the next section. Aim of this report is to outline how the DMP can be applied to the biodiversity FDOs detailed in openDS including classes for DES, human and machine agents or media objects containing image or audio streams (cp. color coding in Figure 3).

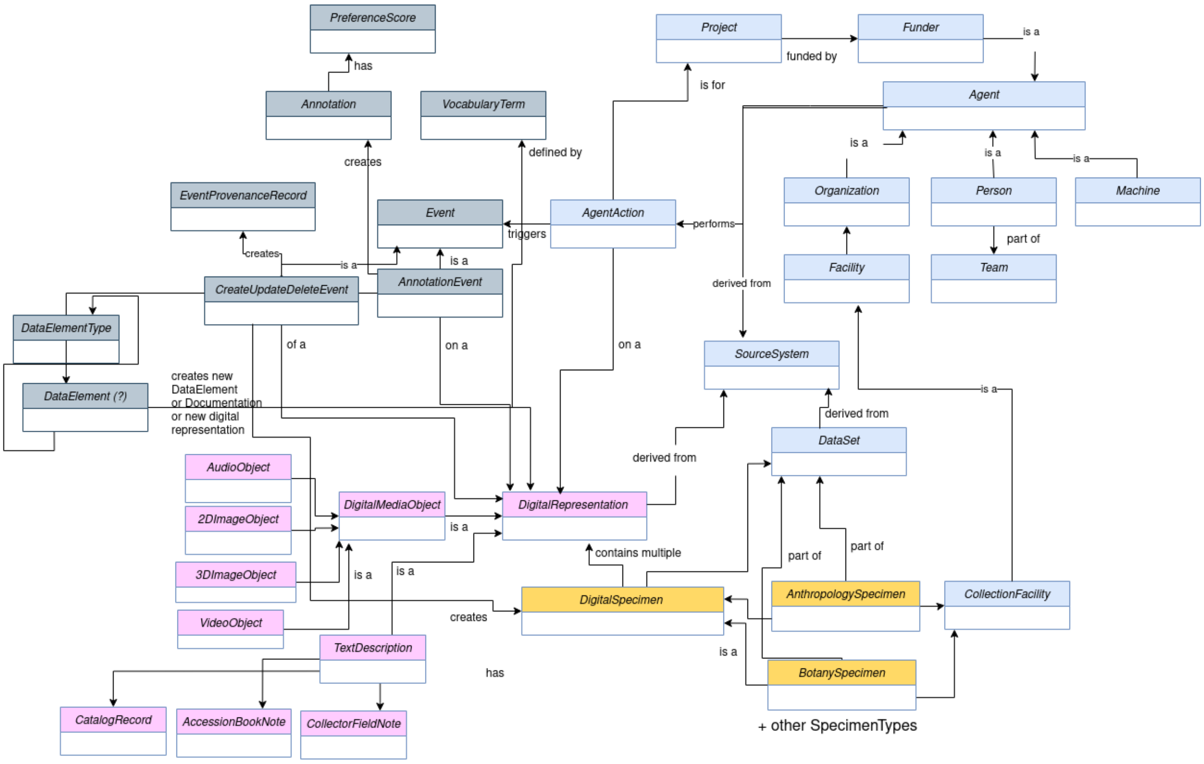


Figure 3: Part of the openDS draft data model showing the core classes for DES like DigitalSpecimen and subclasses (yellow), Agents like Person and Machine (light blue), DigitalMediaObjects like

2DImageObject and AudioObject (pink) and the classes for annotations and provenance like AnnotationEvent and EventProvenanceRecord (dark blue).

The Provenance Ontology (PROV-O) used in openDS is part of a family of provenance related specifications released by W3C, that share a common conceptual data model (PROV-DM). The data model, illustrated in Figure 4, consists of a set of core types (Entity, Activity, Agent) and relations between these types (shown in Figure 4 are those for generation, usage, derivation, attribution and association). The rationale behind this is to foster interoperable exchange of provenance data between systems which use separate domain and application specific descriptions for provenance by provision of a common data model as mapping target (we will see this mapping pattern more frequently in the course of this report).
Provenance in PROV-O describes the provenance of physical, conceptual or digital items e.g. a DES as an instance of the prov:Entity class. A prov:Activity describes a process linked to generation and usage of those entities like a digitisation activity which creates a DES. In most cases, a revision or an update of an entity is triggered by usage of an activity.
A prov:Agent is associated with an activity so that the agent - a human, a software agent, a non-human legal person etc. - is accountable for the process, for instance a machine learning service operating an enrichment activity. Since those non-human agents act on behalf of others, e.g. a machine learning pipeline on behalf of a curator responsible for specimen, PROV-O enables the representation of such chains of accountability.

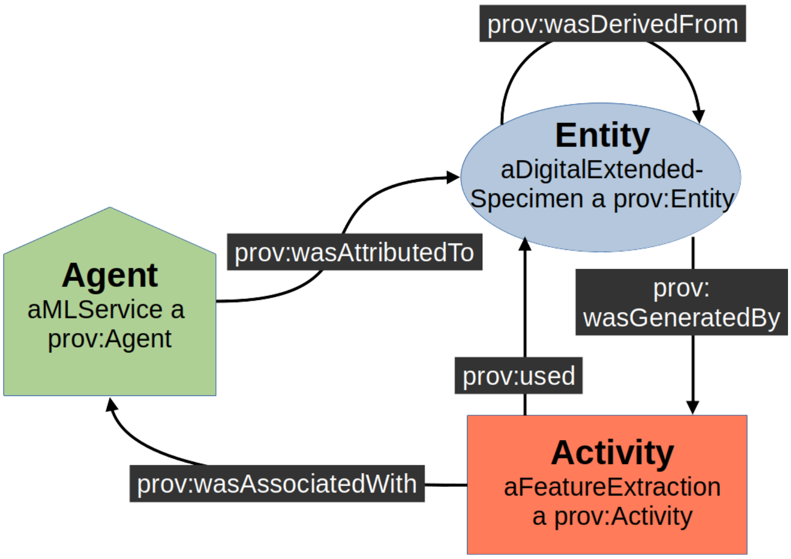


Figure 4: Standard W3C Provenance Data Model linking the basal classes Entity, Activity and Agent through relations generation, usage, attribution and association. Revisions and updates of an entity are special cases of derivation where the revised entity is a version of some original by utilization of an entity by an activity (activity → entity → entity').

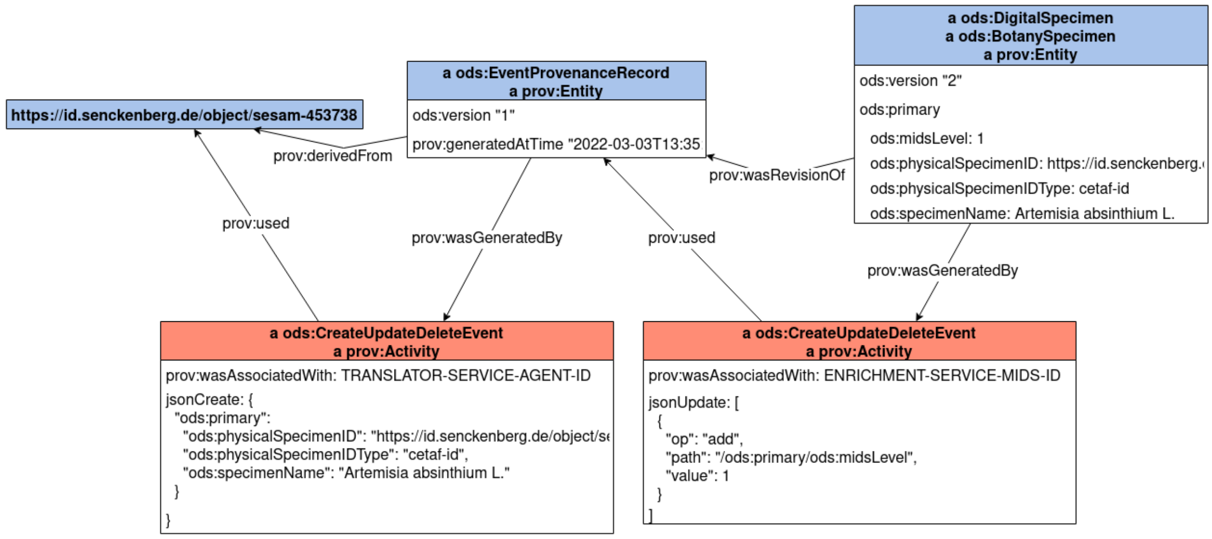
The following implementation section details how PROV is combined with the JSON Patch format (RFC6902) to prototype the operation semantics for prospective provenance to perform operations on a DES involving incremental additions, removals, or substitutions of elements and attributes.

Implementation of provenance recording in DiSSCo as machine actionable DMP

Within the framework of openDS, specific classes and properties for the comprehensive provision of provenance data were designed and developed (e.g. ods:EventProvenanceRecord).
Based on these classes, several proto-typical use cases involving different agents and roles like machine-based annotation, term/taxonomic revision by curator, enrichment by scientist were implemented.
Subsequently, serializations to capture retrospective provenance data including the details of executed processes to derive a specific result were developed involving standards like PROV-O and the Web Annotation Ontology.

A major challenge is the implementation of a provenance recording mechanism which accounts for traceability as well as restartability. The PROV-DM provides the semantic structure to represent the activities and agents involved to come to a certain state of an entity (here the DES). This means that by recording the history of operations on a DES with the PROV vocabulary, the requirement of traceability is fulfilled since human or machine agents can trace a DES in the DiSSCo infrastructure back to its origins, e.g. to the point when it was first created based on a record from a museum's CMS.

However, since the PROV-DM is domain-agnostic (Moreau et al. 2013), it does not include a vocabulary needed to fulfill the requirement of restartability of a DES. The lifecycle of a DES as a digital object consists of create, update and delete events, which are represented in the CreateUpdateDeleteEvent class (Figure 3).


a ods:DigitalSpecimen
a ods:BotanySpecimen
a prov:Entity
ods:version "2"
ods:primary
ods:midsLevel: 1
ods:physicalSpecimenID: https://id.senckenberg.
ods:physicalSpecimenIDType: cetaf-id
ods:specimenName: Artemisia absinthium L.
a ods:EventProvenanceRecord
a prov:Entity
ods:version "1"
prov:generatedAtTime "2022-03-03T13:35
https://id.senckenberg.de/object/sesam-453738
prov:derivedFrom
prov:wasRevisionOf
prov:used
prov:wasGeneratedBy
prov:used
prov:wasGeneratedBy
a ods:CreateUpdateDeleteEvent
a prov:Activity
prov:wasAssociatedWith: TRANSLATOR-SERVICE-AGENT-ID
a ods:CreateUpdateDeleteEvent
a prov:Activity
prov:wasAssociatedWith: ENRICHMENT-SERVICE-MIDS-ID

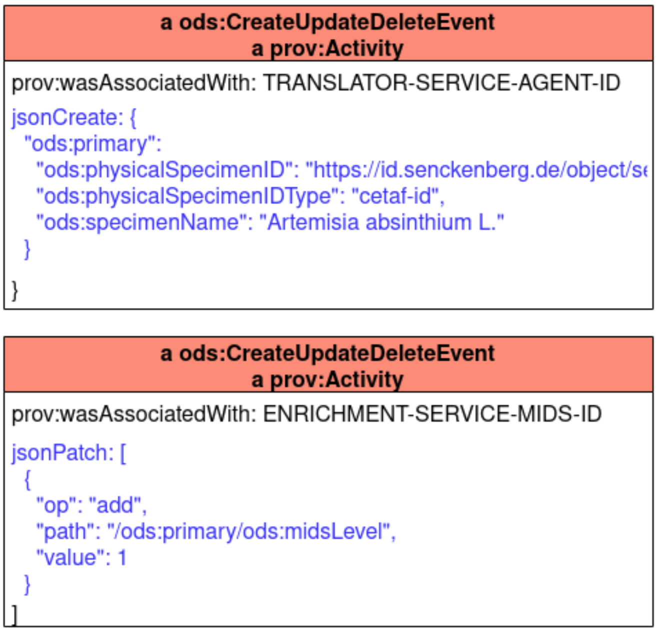
```
jsonCreate: {
  "ods:primary":
    "ods:physicalSpecimenID": "https://id.senckenberg.de/object/se
    "ods:physicalSpecimenIDType": "cetaf-id",
    "ods:specimenName": "Artemisia absinthium L."
  }
}
```

```
jsonUpdate: [
  {
    "op": "add",
    "path": "/ods:primary/ods:midsLevel",
    "value": 1
  }
]
```

Figure 5: A set of nodes representing the provenance graph for a Digital Specimen in alignment with the W3C PROV-DM.

In order to account for restartability it is therefore required to store the exact operations which are performed on the DigitalSpecimen class during a CreateUpdateDeleteEvent to transform it from one state into another. By additionally recording the timestamp of the event it is then possible to restore a previous state of a DES of any point in time by recursively applying the operations from the initial state of the DES when it was created onwards. Several technologies allow the calculation of such patch transformations based on the comparison (difference) between two states of a digital artifact, depending on its data format. Since the data exchange format for DES in the DiSSCo infrastructure is JSON (Glöckler 2022, Leeflang 2022), we can make use of a library which implements the JSON Patch standard. Upon the update event of a DES which contains its new state, the patch transformation is calculated based on the DES's previous state and is persisted in its provenance as part of a new CreateUpdateDeleteEvent instance (Figure 5).

To address the requirement for prospective provenance an approach combining JSON Patch and PROV was drafted proving a recipe of the computational steps to reproduce particular results of operations on a DES making the DES restartable in the current or a former version (Figure 6).



a ods:CreateUpdateDeleteEvent
a prov:Activity

prov:wasAssociatedWith: TRANSLATOR-SERVICE-AGENT-ID

```
jsonCreate: {
  "ods:primary":
    "ods:physicalSpecimenID": "https://id.senckenberg.de/object/s
    "ods:physicalSpecimenIDType": "cetaf-id",
    "ods:specimenName": "Artemisia absinthium L."
  }

}
```

a ods:CreateUpdateDeleteEvent
a prov:Activity

prov:wasAssociatedWith: ENRICHMENT-SERVICE-MIDS-ID

```
jsonPatch: [
  {
    "op": "add",
    "path": "/ods:primary/ods:midsLevel",
    "value": 1
  }
]
```

Figure 6: Detailed view of the Activity nodes in Figure 5. While the general provenance graph can be semantically interpreted, the part highlighted in blue is non-semantic, plain JSON notation. The value of the key "jsonPatch" (bottom) is compliant with RFC6902 and can transform the JSON value of the key "jsonCreate" (top) into another state (restartability)

The implementation of DMPs as FDOs will enable the embedding of DMPs in existing workflows and consequently make the automatic generation of comprehensive reports as well as provenance records possible.

The concept of a machine-actionable DMP (maDMP) aims at addressing some of these issues by making the DMP machine-actionable and enabling the creation of downscaled think reports containing structured data by keeping the human-readability of these reports. The adoption of an open, shared and interoperable concept of maDMP could bring multiple benefits, such as facilitating data discovery and reuse, and enabling automated evaluation and monitoring.
Similarly, information from DMPs can be used to trigger actions, for example, the license and embargo selected by a researcher can be used to automatically fill out information on data deposited into a repository.

Enable linking to the wider context of European and global data federations

Considerable efforts have been undertaken to embed the DES concept in the wider biodiversity service landscape, in particular with regard to interoperability with external taxonomic and geo-collections services (Woodburn 2022) and the alignment of openDS with new extensions of GBIF's core data model to improve coverage, description and presentation of diversified data types including environmental DNA (eDNA), between- and within-species interaction networks or taxonomic treatments (Robertson 2022).

Major challenges for cross-domain reuse of data beyond those community efforts are the agreement on common data standards as well as setting up the corresponding information infrastructures for data exchange between the different research fields involved (Collins 2018).
As detailed before, the concept of FDOs and the related DOA is considered within DiSSCo as the most promising overarching approach to realize an integrated data space of autonomously interacting domain-specific knowledge units (de Smedt 2020). As pointed out above, FDOs solve fundamental problems with regard to machine-actionability by enabling machines to autonomously discover registered instruction sets and process content provided by a FDO (Lannon 2021). But the capability of machines to handle the data encapsulated in those objects semantically appropriate - to make "Machines know what it means" (Mons 2020) - would still require either harmonization or extensive mappings of the various schemas and profiles employed by data providers and communities from different domains in shared data spaces. FDOs couldn't establish and maintain cross-domain interoperability if an intractable number of provider and community-based schemas requires alignment to enable self-contained processing of the described data by machines.

This basic problem of cross-domain interoperability affects DiSSCo in particular with regard to integration in hyper infrastructures like EOSC and the ENRVI subcluster. Essentially two strands to achieve interoperability on data and service level emerge:

- ○ The provision of a universal specification like Cross-Domain Interoperability Framework (CDIF) established in the context of the Global Open Science Cloud, which combines elements of, among others, Schema.org, DCAT, SKOS, PROV-O and I-ADOPT (Gregory 2022). Involved infrastructures align their domain vocabulary with one unified standard provided by the framework.

- The FDO approach to achieve interoperability of self-contained exploration and processing of an FDO's structured data by machines across domains. This capability is based on the interplay of retrievable overarching and machina-actionable descriptions in form of schemas, central (i.e. organized by the FDO Forum) and community based registries for schemas and operations, and the PID system. This meshing of descriptions (retrievable attribute/value pairs) and infrastructure enables machines to obtain information about what they can expect in an FDO and how to operate on it (details in Figure 7).

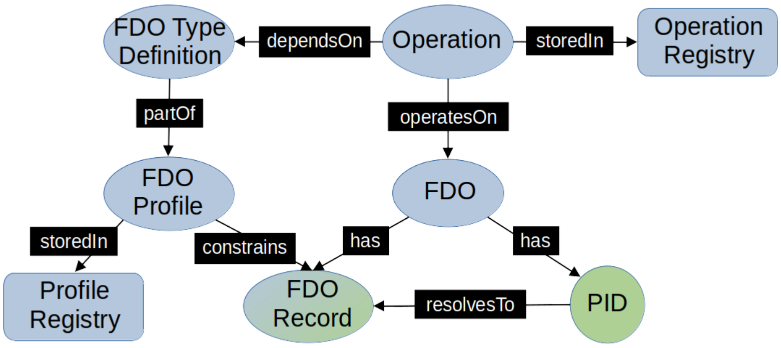


Figure 7: Entry points for this autonomous exploration are FDO records providing attribute-value pairs offered to a machine when resolving the FDO's PID. An additional FDO profile also constrains the structure of the FDO record and provides a description that machines know what they can expect in the record and how to operate on it.

The entry point for a machine to explore an FDO is the FDO record which is obtained by resolving the FDO's PID. This record provides machine-readable attribute/value pairs, it is therefore implemented as FDO itself. The set of expectable and mandatory attributes is constrained by a so-called profile. Profiles themselves are FDOs and rely on specific profile registries (Figure 7).

It becomes clear that at the current state the FDO approach focuses on the interplay of FDO records, profiles, PID system and operations plus their underlying technical infrastructure (in particular registries). Attribute sets e.g. employed in specify profiles, will have to follow a clear syntax and semantics but they are yet not constrained by any semantic artifact and defined in a particular ontology - efforts to design overarching "cross-domain" standards are in this respect rather complementary than contrarily.

DiSSCo has to pursue both approaches simultaneously (while elaborating their interdependencies) depending on the particular project context:

- In the context of the ENVRI-FAIR project, the cluster of European Research Infrastructures (ENVRI) including DiSSCo builds a network of interoperable FAIR data services with the major objective connect these data services to EOSC and promote the reuse of provided datasets based on extensions of the DCAT-AP for the EPOS Research Infrastructure.

- Within the Horizon Europe Project *Biodiversity Digital Twin for advanced simulation, modelling and simulation capabilities* (BioDT) DiSSCo will foster the FDO approach together with partners like LifeWatch ERIC and GBIF to provide a core data model to develop a Digital Twin bringing machine-actionability and FAIR-compliance to high-performance computing (HPC) to address grand challenges in Biosphere Research.

We proposed and prototypically implemented a schema based on W3C Data Catalog Vocabulary (DCAT)[1] to facilitate the mapping of core entities of DISSCo like ods:DigitalSpecimen and ods:DigitalSpecimenCollection. Since openDS comprises a specification for the core data models, but doesn't aim to characterize the related infrastructures, these - like CMSs - will be directly described in DCAT terms.

DCAT is a widely used RDF vocabulary provided by W3C to enable interoperability between data catalogs on the Web. Comprehensive alignments exist to schema.org - the other lingua franca for the web - and domain vocabularies like INSPIRE 19139. Based on the fundamental vocabulary, a set of application profiles (DCAT-APs) were proposed, which provide the metadata records to meet the specific application needs of data portals (mostly) in Europe. Official EU services use the DCAT-AP while the European Plate Observing System (EPOS), developed its own AP which will be reused in the ENVRI-FAIR network.

The essential elements of DCAT are centered around the representation of a collection of data on the web (Figure 1):

- dcat:Catalog describes a curated collection of metadata about datasets
- dcat:Dataset outlines collection of data, published or curated by an agent
- dcat:CatalogRecord provides the metadata of a dcat:Dataset
- dcat:Distribution defines specific available form of a dataset (actual file format)
- dcat:Resource is overarching class for all things published or curated by an agent
- dcat:DataService represents collection of operations that provides access to one or more datasets or processing functions

DCAT elements are descriptions of files, the file content itself is either stored in a file system or provided by a service.

[1] https://www.w3.org/TR/vocab-dcat-2/

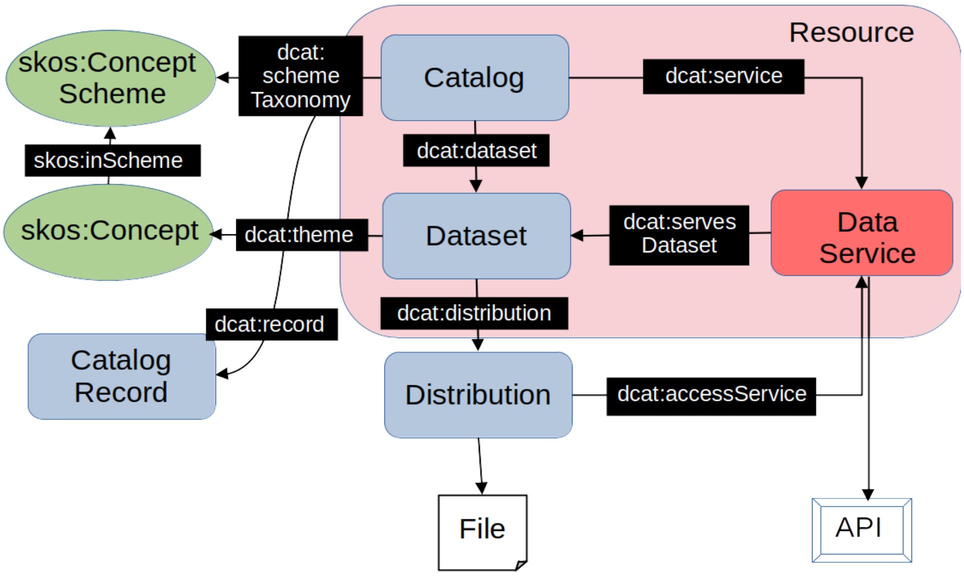


Figure 7: Essential components of DCAT to represent data catalogs on the web. Blue elements were part of DCAT1 specification, red elements were introduced with the extended schema DCAT2.

Figure 8 demonstrates a schema developed for Earth System Sciences which comprises many entities reusable to provide mappings for NSCs (dcat:Catalog) and affiliated services (dcat:DataService) from openDS to DCAT.

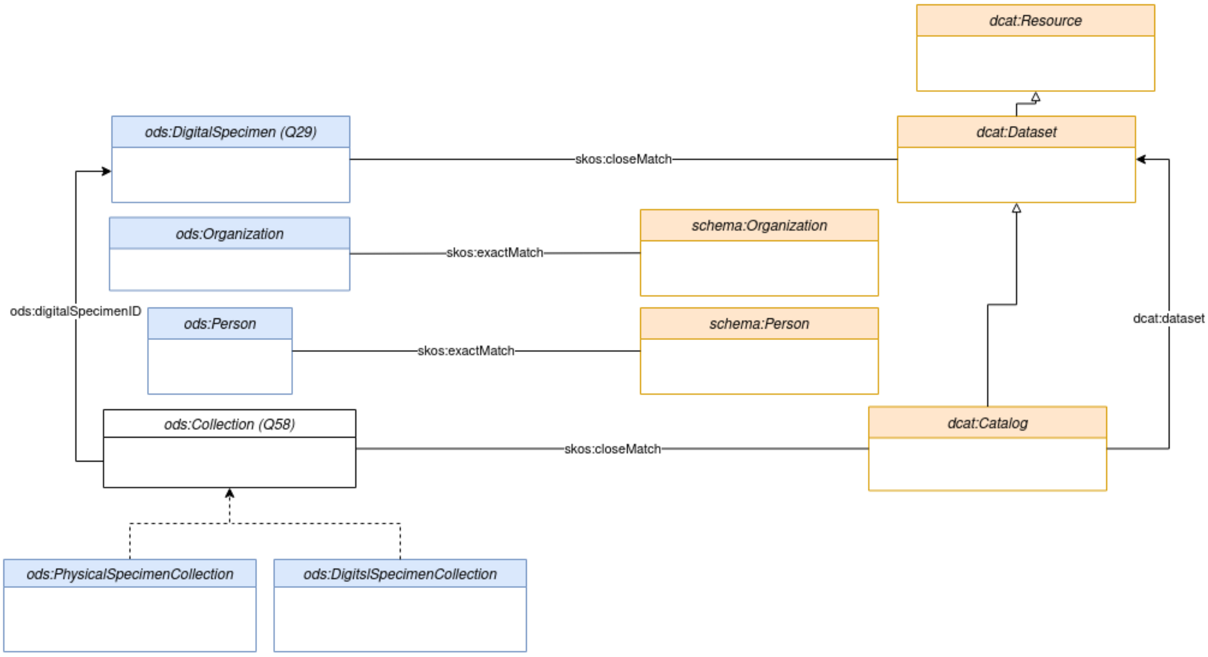


Figure 8: Possible mapping between some ODS classes (blue) and their matching classes included in the EPOS DCAT application profile (orange).

To sum up, we will now analyze potential gaps in DiSSCo's maDMP/FDO approach regarding interoperability and reusability of data with the EOSC. Realizing both technical and semantic interoperability within EOSC is crucial for the emerging data and service federation that will constitute EOSC. For this purpose, essential guidelines concerning this matter where compiled into the EOSC interoperability framework (EOSC IF; European Commission 2021)

In Table 1 the current status achieved for data management in DiSSCo is aligned with the EOSC IF recommendations for the technical and semantic layer as checklist.

EOSC IF recommendation	DiSSCo implementation	Comment/Xref
Open Specifications for EOSC Services	FDO specifications are provided by the FDO Forum. DiSSCo architecture is detailed in DPP D6.2 Implementation and construction plan of the DiSSCo core architecture. API guidelines derived from the existing specifications of JSON:API, OpenAPI and CloudEvents.	DiSSCo provides comprehensive open specifications for the DiSSCo eServices. Leeflang 2022 Glöckner 2022 Islam 2020
A common security and privacy framework (including AAI infrastructure)	FAIR Digital Objects can implement trust and privacy mechanisms, though currently all DES data is provided as open data.	Roles are implemented in the central AAI system (Keycloak), which can be implemented in the FDO layer based on ODRL (DES data published at this stage is FAIR/O).
Easy-to-understand Service-Level Agreements for all EOSC resource providers	This is work in progress and will be part of DiSSCo Create.	To be implemented in DiSSCo Create.
Easy access to data sources available in different formats	DwC/IPT ABCD/BioCASE CETAF-ID via http	See D 6.1. Glöckner 2022
Coarse-grained and fine-grained dataset (and other research object) search tools.	Closely related to maDMP FDO (granularity of output).	To be implemented in DiSSCo Create.
A clear EOSC PID policy	DiSSCo adopted a DOI-driven approach for the persistent identification of DES.	Driven-by-DOI Hardisty 2021

Clear and precise, publicly-available definitions for all concepts, metadata and data schemas	The openDS github repository details all elements of the openDS specification.	Documentation of the specification on github (https://github.com/DiSSCo/openDS), additional documentation of terms in the DMF Fichtmüller 2022 DMF https://modelling.dissco.tech
Semantic artifacts preferably with open licenses.	DiSSCo's Technical Team released a recommendation which encourages the use of permissive licenses (e.g., MIT or Apache V2.0).	Not adopted yet for semantic artifacts in DiSSCo.
Associated documentation for semantic artifacts.	Comprehensive documentation for the openDS on github	Inconsistencies between documentation in DMF and github (https://github.com/DiSSCo/openDS). For DES see: Hardisty 2022
Repositories of semantic artifacts, rules with a clear governance framework	DMF comprises Wikibase as an open development platform.	Governance framework nor clear yet
A minimum metadata model (and crosswalks) to ease discovery over existing federated research data and metadata.	Metadata models provided in the openDS data model.	Elementary bindings to domain standards DwC and ABCD, started mapping to DCAT, Islam 2020
Extensibility options to allow for disciplinary metadata.	DES as an FDO Type inherits the vast extensibility options of the FDO approach of abstraction and encapsulation.	Biodiversity FDO Types are easily extensible within the FDO Typing framework.
Clear protocols and building blocks for the federation/harvesting of semantic artifacts catalogs.	Biodiversity FDO Types like ods:DigitalSpecimenCollection can be harvested from Digital Object Repositories via DOIP and HTTP REST. DiSSCo's core architecture uses opens community protocol to harvest data from participating providers (IPT/DwC & BioCASe/ABCD)	Data harvesting in DiSSCo builds upon well established standards biodiversity and web community standards.

Table 1; Compliance of DiSSCo's data services and specifications with EOSC IF recommendations w.r.t to technical and semantic interoperability.

Excursus: Storage infrastructure

DiSSCo will open up an increasing amount of genomic, environmental & phenotype data from participating NSCs and will foster in this way further large-scale digitization efforts producing vast volumes of data. These data will include two-/three-dimensional digital images of the specimen plus accompanying images like computerized tomography scans, converted labels and markings on the specimen into machine-readable text as well as the digital recording of discovery sites and habitats.
Major challenges are the development of the information infrastructures that provide methods for integrating these data and metadata to provide easy data exchange among researchers, institutions and services (a first assessment with regard to long-term storage and archiving of specimen data and assigned multimedia files has been conducted in the mobilse Workshop *Data storage and archiving strategies*, Kahila Bar-Gal 2019).

As outlined before, a DES is an abstract logical structure of information content, which is or might be located to large extent in remote repositories (DES contains essentially links), not locally (DES encapsulates only core data as bitstream).

However, storage requirements of digital objects depend extremely on purpose and aspired quality, where DiSSCo expects digital images as the main source for storage requirements (Table 2).

Bare Digital Extended Specimen (no bitstream of images or sequences)	~ 2-10 KB
Full genomes	10 MB (fungi) >1GB (plants, lungfishes)
Herbarium Scans	10MB - 200MB (?) (https://doi.org/10.3897/rio.6.e50675)
Insect tray (~30 insects)	500MB (10.3897/zookeys.209.3178)

Table 2: Comparison of storage requirements of different digital assets.

The DiSSCo Data Storage container is part of DiSSCo's core architecture and accordingly detailed in the construction master plan. Its main aim is to provide a storage component for the processing pipeline - data is continuously accessed here, e.g. for retrieval and indexing.

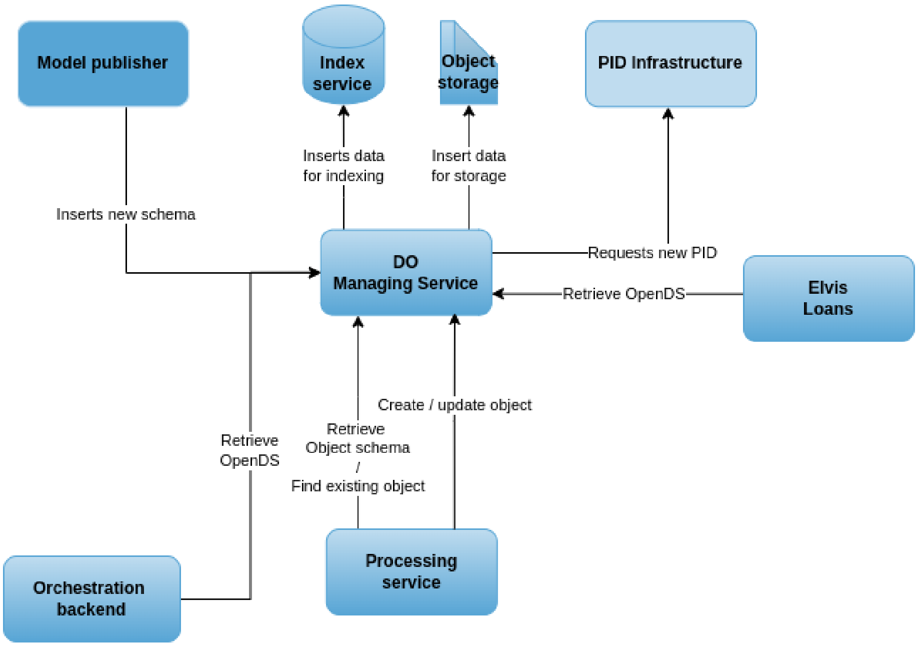


Figure 10: Component diagram of the DiSSCo digital object managing service including the object storage (reused from Leeflang 2022).

However, a so-called persistence store is planned to store data for longer periods. Both the metadata (digital object) and the data (images, sounds, etc.) may be stored.

From the *Implementation and construction plan of the DiSSCo core architecture* (Leeflang 2022):

"...Potentially the amount of data could grow to large dimensions which means that the storage solution needs to be able to grow with the data. Besides storage capacity the uses of the data will also grow, meaning the solution needs to be scalable based on the incoming requests.
"

According to Leeflang et al., main objective of the data storage in the core architecture is to support DISSCo's e-services for collaborative annotation, curation ("UCAS") and knowledge discovery centered around the DES - it will not provide generic storage for bulk digitization output in arbitrary quality and resolution.

However, DiSSCo could provide blueprints for the set-up of storage facilities in compliance with the central service architecture. Aim of a local storage center will be to ingest, enrich, and integrate digitized data into a local digital object repo and to correspond with the DiSSCo core facilities . Rich metadata involving the MIDS specification and provenance data should be applied to the digitized object. Multi-Petabyte capacity for accompanying digital images can be provided based on an open clustered object storage platforms, but their implementation is currently not foreseen as task for the DiSSCo Hub Desirable is the integration and embedding in a virtualisation and abstraction layer consistent with current standards like the FAIR Digital Object approach.

Discussion and Outlook

The theoretical core of the Digital Extended Specimen concept – DES as a specific type of FDO that acts as a digital surrogate for a specific physical specimen in a natural science collection (Islam 2022) - was significantly elaborated with the term of DiSSCo Prepare. Beyond the further adoption of the concept in the biodiversity community, driven by DiSSCo and the US-based Biodiversity Collections Network (BCoN, Ellwood 2021, Hardisty 2022), the approach is now also an integrated part of the EU's *Destination Earth* (DestinE) initiative is to develop an accurate digital model of the Earth via sets of thematic Digital Twins. DiSSCo and partners will develop and extend the DES as a core data model for the biodiversity Digital Twins, which will bridge between advanced simulation using biodiversity Big Data (La Salle 2016) on HPC platforms and the FAIRification of these data, employed computational/laboratory methods and workflows.
Snapshotting of states - called Checkpoint/Restart approach (C/R; Zhao 2021), is essential for a variety of tasks in the context of HPC such as fault-tolerant computing, flexible scheduling of compute resources, migration of computations between systems and debugging at large scale. The requirements of a biodiversity Digital Twin regarding C/R are to some extent addressed by the prospective provenance prototype for the DES described previously: Based on PROV-O and the JSON Patch, restartability and partial updates for DESs from selectable checkpoints can be realized. The approach is currently further expanded to enable the transfer of operation sequences and workflow parameters to data objects with similar constraints, e.g. the sharing of digitization pipelines between similar subtypes of the DES and which be elaborated to support full workflow software environment preservation involving tools like CWLProv and RO-Crate (Khan 2019, Soiland-Reyes 2022)

The next step towards a machine actionable data management in a FDO ecosystem is now to develop or include if available key services involving automatic recommender services for licensing ike the EUDAT license tool, tools to compose access and update policies using machine-actionable policy elements based on machine-actionable formats like PROV-O and ODRL and services to enable the automatic forwarding of RDM tasks to curators and other agents when specific input is expected.

Acronyms, terms, and definitions

Acronym	Term	Definition in the DiSSCO context
ABCD	Access to Biological Collection Data	Comprehensive standard for access to and exchange of data about specimens and observations.
BCoN	Biodiversity Collections Network	Initiative funded by the U.S. National

		Science Foundation to foster digital availability of all U.S. biodiversity collections for research, education, decision-making, and other scholarly and creative activities.
BioCASe	Biological Collection Access Service	Meant in this context is the BioCASe Provider Software, a data binding middleware that allows publishing of multiple data resources of a provider with a single web service in ABCD format.
BioDT	Biodiversity Digital Twin for advanced simulation, modelling and simulation capabilities	Large EU-funded project employing FAIR data combined with digital infrastructure, predictive modeling and AI solutions to implement a Biodiversity Digital Twin, which aims to facilitate the development of evidence-based solutions for biodiversity protection and restoration.
CD	Collection Descriptions	Data standard for describing collections of natural history materials including information about access and usage of specimens.
	Dataset	A collection of data, published or curated by a single agent, and available for access or download in one or more representations.
CDIF	The Cross-Domain Interoperability Framework	Framework supporting cross-domain interoperability by development of a (set of) common standard(s) which enables a “one-to-many” mapping against the multitude of data and service providers domain standards.
DCAT	Data Catalog Vocabulary	W3C (RDF) vocabulary to foster interoperability of data catalogs on the web based on a unified description of data resources (datasets and data services).
DCAT-AP	DCAT Application Profile for data portals in Europe	Standardized profile built upon DCAT to enable a uniform description of public sector datasets in Europe.
DES	Digital Extended Specimen	Digital Twin of a physical specimen in a collection, encapsulates and persistently links to information artifacts derived from the physical specimen such as sequences, images and taxonomic determinations.
DestinE	Destination Earth	Framework within the European Commission’s Green Deal and Digital

		Strategy with the objective to develop thematic digital earth twins as basis for an accurate digital model of the Earth to monitor and predict the interaction between natural phenomena and human activities.
DMP	Data Management Plan	Formal document detailing how data is handled during/after a (research) project.
DOA	Digital Object Architecture	DOA introduces the concept of a digital object, which forms the basis for the architecture by specification of three key components: Identifier/resolution system, repository system, and registry system.
DMF	DiSSCo Modelling Framework	An instance of Wikibase where the DiSSCo data model is being developed.
DwC	Darwin Core	Framework of standards to compile and mobilize biodiversity data from varied and variable sources.
ENVRI	Environmental Research Infrastructures	Organizational integration of European Research Infrastructures with a focus on Earth system research including amongst others DiSSCo, LifeWatch and EPOS.
ENVRI-FAIR		Initiative of the ENVRI Cluster to build FAIR services for research, innovation and to connect them to the emerging EOSC.
EOSC	European Open Science Cloud	Emerging cross-disciplinary infrastructure to foster seamless and FAIR-compliant access to and reuse of European research data.
EOSC IF	EOSC Interoperability Framework	Set of guidelines and services (registries) that promote standards and community best practices within EOSC to achieve interoperability of data and services
EUDAT CDI	EUDAT Collaborative Data Infrastructure	Large distributed infrastructure of integrated data services and computing resources for research in Europe. Started as *European Data Infrastructure* project and is now sustained by a network of more than 20 European research centers as EUDAT CDI.
EPOS	European Plate Observing System	European Research Infrastructure Consortium with the objective to create single, Pan–European, sustainable and distributed infrastructure for solid Earth science.

FAIR	The FAIR Data principles	Guidelines to improve and foster reuse of digital research assets with respect to the four foundational principles Findability, Accessibility, Interoperability, and Reusability.
	FAIRification	Process describing the stepwise transformation from an initially non-FAIR dataset to deployment as FAIR data resource.
FAIR/O	FAIR and Open Data	FAIR/O is a new introduced term (Quay 2022) to describe that a dataset or a data infrastructure complies with both FAIR and Open Data principles and standards.
FDO	FAIR Digital Object	FDOs are abstracted data objects encapsulating content, descriptive metadata and globally resolvable and persistent identifiers in compliance with the FAIR principles.
GBIF	Global Biodiversity Information Facility	Networked data infrastructure funded by the world's governments aggregating data about all types of life on Earth.
ICEDIG	Innovation and Consolidation for Large Scale Digitisation of Natural Heritage	Project which provided essential blueprints and capacity enhancements to make DiSSCo operational with special emphasis on mass digitisation and subsequent access to all related data
IPT	GBIF Integrated Publishing Toolkit	A free, open source software tool used to publish and share biodiversity datasets through the GBIF network
I-ADOPT	InteroperAble Descriptions of Observable Property Terminology	RDA ontology framework developed to enable interoperability between existing variable descriptions in semantic artifacts like ontologies, taxonomies, and controlled vocabularies.
INSPIRE	Infrastructure for Spatial Information in Europe	Geodata infrastructure instigated by the European Commission to enable the sharing of environmental spatial data between public sector organizations and provide public access to these data.
JSON	Java Script Object Notation	Lightweight, text-based, language-independent interchange format for structured data.
JSON-LD	JSON for Linking Data	JSON-LD is a syntax to serialize Linked Data in JSON and can therefore be used as an RDF syntax.

LD	Linked Data	Set of methods to publish structured and connected data involving i.a. controlled vocabularies and ontologies to enable machine-interpretability of data.
LifeWatch	LifeWatch ERIC	European Research Infrastructure Consortium setting up e-Science infrastructures like EcoPortal for ecology semantics to support research of biodiversity and ecosystem functions and services.
maDMP	Machine actionable Data Manegement Plan	Technically advanced DMPs complemented by automatized features for (i.a.) reporting, workflow embedding and policy assignment.
MIDS	Minimum Information about a Digital Specimen	Specification defining information elements for graded digitization levels of physical specimens.
NSC	Natural Science Collection	Term used within DiSSCo to underline the relevance of specimen data for current scientific challenges contrary to a historicist understanding of collection data.
ODRL	Open Digital Rights Language	Rights expression language for the specification of policies and rights to manage digital assets, i.e. to define such policies, roles and conditions for reuse of data in a machine-readable way.
openDS	Open Digital Specimen	Specification providing the set of elements to model DES (and other curated biodiversity objects) as typed data objects compliant with FDO specifications.
PROV-DM	PROV Data Model	Generic data model for provenance underlying the PROV family of specifications. PROV-DM provides core concepts to foster the interoperability of various separate provenance formats by alignment with a common specification.
PROV-O	PROV Ontology	PROV-O provides a set of classes, properties and relations to represent provenance information in a variety of application domains.
RO-Crate	Research Object Crate	Lightweight approach to provide digital object containers for research data together with related metadata using on schema.org annotations.
SKOS	Simple Knowledge Organization	Formal framework for the development of

	System	knowledge organization systems like thesauri and controlled vocabularies for the Semantic Web.
RDA	Research Data Alliance	Large initiative driven by the global research community to develop the social and technical bridges that enable open sharing and re-use of data.
RDF	Resource Description Framework	Standard data model of the Semantic Web to model resources and statements about those as triples in a knowledge graph.
	Schema.org	Comprehensive widely spread structured data vocabulary for web services with the aim to improve the findability of resources on the web.
W3C	World Wide Web Consortium	W3C is an international community which develops and issues essential Web standards including HTML5, the Semantic Web stack, XML, and SPARQL.

References

Addink W, Hardisty AR (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

Alliance for Biodiversity Knowledge (2021) Converging Digital and Extended Specimens: Towards a global specification for data integration, https://www.allianceforbio.org/post/converging-digital-and-extended-specimens-towards-a-global-specification-for-data-integration (retrieved 2022-12-16)

Ashmore, R., Calinescu, R., Paterson, C.. 2021. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. ACM Comput. Surv. 54, 5, Article 111 (June 2022), 39 pages. https://doi.org/10.1145/3453444

Belhajjame, K., B’Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J. and Miles, S., 2013. Prov-dm: The prov data model. W3C Recommendation, 14, pp.15-16. https://www.w3.org/TR/prov-dm/ (Retrieved on 5 December 2022).

Cardoso, J., Proença, D., Borbinha, J. (2020). Machine-Actionable Data Management Plans: A Knowledge Retrieval Approach to Automate the Assessment of Funders' Requirements. In: , et al. Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science(), vol 12036. Springer, Cham. https://doi.org/10.1007/978-3-030-45442-5_15

Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Laaksonen, L., Mietchen, D., Petrauskaitė, R.. Wittenburg, P., 2018, Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data https://doi.org/10.2777/54599

Davidson, S. and Freire, J. 2008. Provenance and scientific workflows: challenges and opportunities. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08). Association for Computing Machinery, New York, NY, USA, 1345–1350. https://doi.org/10.1145/1376616.1376772

de Natale, Flora, Bossi, Monique, Chanzy, Andre, de Nart, Dario, de Pascalis, Francesca, Ferrighi, Lara, Fiore, Nicola, Gasner, Paul, García Rodríguez, J. M., González-Aranda, Juan Miguel, Islam, Sharif, Ignatiuk, Dariusz, L'Abate, Giovanni, Parisse, Barbara, Peterseil, Johannes, Pichot, Christian, Rosati, Ilaria, Sáenz-Albanés, A. J., Sánchez Cano, F. M., … Basset, Alberto. (2021). ENVRI-FAIR D11.2: Report on FAIRness implementation activities in the Biodiversity and Ecosystem subdomain (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4682826

De Smedt, K., Koureas, D. and Wittenburg, P., 2020. FAIR digital objects for science: from data pieces to actionable knowledge units. Publications, 8(2), p.21. https://doi.org/10.3390/publications8020021

Ellwood, Elizabeth R., Bentley, Andrew, Buschbom, Jutta, Hardisty, Alex, Mast, Austin, Miller, Joe, Monfils, Anna, Nelson, Gil, & Paul, Deborah L. (2021). Highlights and Outcomes of the 2021 Global Community Consultation. Biodiversity Information Science and Standards, 5, e72716. https://doi.org/10.3897/biss.5.72716

European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., et al., EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture, Publications Office, 2021, https://data.europa.eu/doi/10.2777/620649

Fichtmueller D. & Güntsch A. (2022) DiSSCo Prepare Deliverable 5.2 "DiSSCo Modelling Framework". https://doi.org/10.34960/e3nv-zh69

Glöckler F., Pim Reis J., von Mering S., Petersen, M., Weiland, C., Dillen, M., Leeflang, S., Haston, E., Addink, W., Fichtmüller, D. (2022), DiSSCo Prepare report D6.1 Harmonization and migration plan for the integration of CMSs into the coherent DiSSCo Research Infrastructure, https://know.dissco.eu/handle/item/490

Grieb J, Weiland C, Hardisty A, Addink W, Islam S, Younis S, Schmidt M (2021) Machine Learning as a Service for DiSSCo's Digital Specimen Architecture. Biodiversity Information Science and Standards 5: e75634. https://doi.org/10.3897/biss.5.75634

Gruber, T. (2009). Ontology. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_1318

Hardisty, Alex. (2019). Provisional Data Management Plan for DiSSCo infrastructure. Deliverable D6.6. ICEDIG. https://doi.org/10.5281/zenodo.3532937

Hardisty AR, Ma K, Nelson G, Fortes JAB (2019b) 'openDS' – A New Standard for Digital Specimens and Other Natural Science Digital Object Types. Biodiversity Information Science and Standards 3: e37033. https://doi.org/10.3897/biss.3.37033

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

Hardisty AR, Ellwood ER, Nelson G, Zimkus B, Buschbom J, Addink W, Rabeler RK, Bates J, Bentley A, Fortes HAB, Hansen S, Macklin JA, Mast AR, Miller JT, Monfils AK, Paul DL, Wallis E, Webster M, Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet, BioScience, 2022;, biac060, https://doi.org/10.1093/biosci/biac060

Islam, S., Hardisty, A., Addink, W., Weiland, C. and Glöckler, F., 2020. Incorporating RDA outputs in the design of a European Research Infrastructure for natural science collections. Data Science Journal, 19(50), pp.1-14. https://doi.org/10.5334/dsj-2020-050

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C., Goble, C., Guizzardi, G., Kryger Hansen, K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R., Imming, M., Jeffery, K., Kaliyaperumal, R., Kersloot, M., Kirkpatrick, C., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S., Bonino da Silva Santos, L., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M., Willighagen, E., Wittenburg, P., Roos, M., Mons, M., Schultes, E.; FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2020; 2 (1-2): 10–29. https://doi.org/10.1162/dint_r_00024

Kahila Bar-Gal, G., & Triebel, D. (2019) WG4 Workshop "Data storage and archiving strategies", URL: https://costmobilise.biowikifarm.net/wiki/WG4_Workshop_%22Data_storage_and_archiving_strategies%22_in_Sofia_(NMNHS) (retrieved 2022-12-12).

Kahn R, Wilensky R (2006) A framework for distributed digital object services. Int J Digit Libr 6(2): 115-123. https://doi.org/10.1007/s00799-005-0128-x

Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR. Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. Gigascience. 2019 Nov 1;8(11):giz095. https://doi.org/10.1093/gigascience/giz095. PMID: 31675414; PMCID: PMC6824458

Lannom, L. (2021). State of FDO Work SciDataCon 18 Oct 2021 https://drive.google.com/drive/folders/1A4vbovXI0h-x_B66XfI12TzmXaB2Lay1 (retrieved 2022-12-12)

La Salle, J., Williams, KJ. and Moritz, C. (2016) Biodiversity analysis in the digital eraPhil. Trans. R. Soc. B3712015033720150337 http://doi.org/10.1098/rstb.2015.0337

Leeflang, S, Weiland, C, Grieb, J, Dillen, M, Islam, S, Fichtmueller, D, Addink, W, & Haston, E (2022). DiSSCo Prepare D6.2 Implementation and construction plan of the DiSSCo core architecture (1.2). Zenodo. https://doi.org/10.5281/zenodo.6832200

Meineke E., Davies J., Daru B., Davis C. (2019) Biological collections for understanding biodiversity in the Anthropocene Phil. Trans. R. Soc. B3742017038620170386 http://doi.org/10.1098/rstb.2017.0386

Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525

MIDS Task Group (2022) *Minimum Information about a Digital Specimen*. Available from https://github.com/tdwg/mids (retrieved 2022-12-16)

Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLoS Comput Biol 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750

Mons, B., Schultes, E., Liu, F., Jacobsen, A. (2020); The FAIR Principles: First Generation Implementation Choices and Challenges. Data Intelligence 2020; 2 (1-2): 1–9. https://doi.org/10.1162/dint_e_00023

Moreau L, Missier P, Belhajjame K, et al. . PROV-DM: The PROV Data Model. 2013. https://www.w3.org/TR/2013/REC-prov-dm-20130430/ (retrieved 2022-12-16).

Papale, Dario. (2020). ENVRI-FAIR D11.1 Biodiversity and Ecosystem subdomain implementation short term plan (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3885361

Quay, A., Fiske, P., and Mauter, M. (2022) ACS ES&T Engineering 2 (3), 337-346 https://doi.org/10.1021/acsestengg.1c00245

Robertson T, Wieczorek JR, Raymond M (2022) Diversifying the GBIF Data Model. Biodiversity Information Science and Standards 6: e94420. https://doi.org/10.3897/biss.6.94420

Schwardmann, U., 2020. Digital Objects – FAIR Digital Objects: Which Services Are Required?. Data Science Journal, 19(1), p.15. http://doi.org/10.5334/dsj-2020-015

Soiland-Reyes, S., Sefton, P., Crosas, M., Jael Castro, L., Coppens, F., Fernández, J., Garijo, D., Grüning, B., La Rosa, M., Leo, S., Ó Carragáin, E., Portier, M., Trisovic, A., RO-Crate Community, Groth, P., Goble, C. (2022): Packaging research artefacts with RO-Crate. Data Science 5(2) https://doi.org/10.3233/DS-210053

Webster MS, Buschbom J, Hardisty A, Bentley A (2021) The Digital Extended Specimen will Enable New Science and Applications. Biodiversity Information Science and Standards 5: e75736. https://doi.org/10.3897/biss.5.75736

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Wilkinson, M., Sansone, SA., Schultes, E. *et al.* A design framework and exemplar metrics for FAIRness. *Sci Data* **5**, 180118 (2018). https://doi.org/10.1038/sdata.2018.118

Wittenburg P, Strawn G. Revolutions Take Time. Information. (2021); 12(11):472. https://doi.org/10.3390/info12110472

Wittenburg P, Anders I, Blanchi C, Buurman M, Goble C, Grieb J, Hardisty AR, Islam S, Jejkal T, Kálmán T, Kirkpatrick C, Lannom L, Lauer T, Manepalli G, Peters-von Gehlen K Pfeil A, Quick R, van de Sanden M, Schwardmann U, Soiland-Reyes S, Stotzka R, Trautt, Z, Van Uytvanck D, Weiland, C, Wieder P. (2022). FAIR Digital Object Demonstrators 2021 (final). Zenodo. https://doi.org/10.5281/zenodo.5872645

Woodburn M, Addink W, Bánki O, Dijkema T, French L, Glöckler F, Humphries J, Islam S, Leeflang S, von Mering S, Pim Reis J (2022) DiSSCo Prepare Deliverable D5.5 Construction plans for the improvement of technical infrastructure in the areas of geo-collection data and taxonomic services. https://doi.org/10.34960/dzs0-xa94 *(in submission)*

Younis S, Schmidt M, Weiland C, Dressler S, Seeger B, Hickler T (2020) Detection and annotation of plant organs from digitised herbarium scans using deep learning. Biodiversity Data Journal 8: e57090. https://doi.org/10.3897/BDJ.8.e57090

Zhao, Z., & Hartman-Baker, R. (2021). Checkpoint/Restart Vision and Strategies for NERSC's Production Workloads. In Checkpoint/Restart Vision and Strategies for NERSC's Production Workloads. Lawrence Berkeley National Laboratory. http://dx.doi.org/10.2172/1814161 (retrieved from https://escholarship.org/uc/item/48v5r5rj 2022-12-16).