RFC: DiSSCo Kernel and Digital Specimen FDO Record attributes

Authors:	Sharif Islam, Wouter Addink, Soulaine Theocharides
To be reviewed by:	2023-06-30
Revisit Date:	2023-07-14
Status:	Draft | Feedback Requested | **Active** | Abandoned | Superseded
Superseded by:	
Related:	DiSSCo specimen & collection classification

See: RFC Process Description

This RFC supersedes DiSSCo datamodel: DS types & Kernel metadata

Need

To implement FAIR Digital Objects (FDOs), DiSSCo must adopt FDO specifications. This necessitates the use of FDO Profiles, and Object Type definitions to enable machine actionability, which is crucial for the widespread utilisation and processing of digital specimens. FDO Profiles are themselves FDOs that explicitly indicate what metadata attributes may be and should be present in objects with this profile. On the other hand, the object type describes the object that the PID identifies. These attributes can be accessed directly by a machine or service when resolving the PID, eliminating the need to retrieve the full bitstream or payload of the object, which may contain additional metadata and data. This abstraction enables swift and efficient discovery and initial selection of digital objects of interest for further processing. To ensure scalability, the elements in the FDO record should align with the recommendations set forth by the Research Data Alliance (RDA) regarding PID Kernel Information. This means that the metadata should remain relatively static and not require frequent updates. It should consist of a minimal set of essential elements ("kernel") that describe when and by whom the PID was created, its current status, the target of resolution, and the nature of the identified digital object. The type definition specifically describes the digital object itself, including any additional metadata, bit sequences, and available operations on the object. By defining the object type, it becomes easier to understand and manipulate the digital object. Overall, adopting FDO specifications and utilising FDO Records, Profiles, and type definitions enable DiSSCo to effectively implement FAIR principles, allowing for machine-actionable digital specimens on a large scale.

Approach

- We use the ePIC Data Type Registry registry to create and test DiSSCo FDO Profiles. This is a type registry in which machine readable descriptions of both the FDO profiles as well as the type definitions for all digital object types can be created. The FDO profile that describes the FDO can also be based on other profiles (e.g. extending them or reusing attribute definitions, which each get their own PID) and to limit duplication of digital object types, this needs governance. We use this registry rather than creating our own because ePIC plans to develop governance for the process of creating new digital object types by defining their FDO profiles.
- For the profiles and records, we aim to follow the FDO Forum recommendation.
- For a Digital specimen: if there is a physical specimen ID then the PID status follows the status of the physical specimen once the DOI is published, e.g. it may become SPLIT or MERGED.
- Definition of a specimen: A specimen is a curated MaterialEntity or Digital Entity that is preserved or maintained for future observations. It is intended to be a representative of some physical thing as an object of interest to someone, on which observations can be made about the represented object. In DiSSCo it needs to represent a natural object. When it is a MaterialEntity is a thing composed of matter that has some defined boundaries. For example a herbarium sheet with plant material, or a microscopic slide with a thin slice from a rock. A digital specimen is intended to contain all information derived from the specimen so it becomes a surrogate or even digital twin for the specimen.

Relationships between FDO Record and Profile (source: https://doi.org/10.5281/zenodo.7825572)

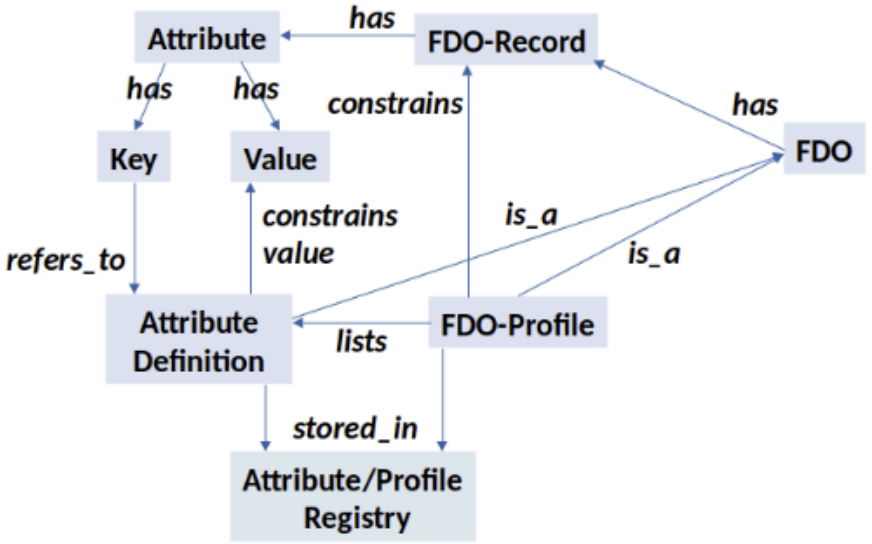

Figure 1: Relationship between FDO Record and Profile

FDO Forum **recommendation:**

For FDO records and their profiles we recommend to use the following set of consistent principles that are derived directly from the RDA KIP principles avoiding inconsistencies. These principles use the FDO glossary and are brought into a new more intuitive order:
FDO-KIP-P1 (from RDA-KIP-P1): The primary purpose of an FDO record (PID Kernel Information record) is to serve machine actionable services.
FDO-KIP-P2 (from RDA-KIP-P6): A profile describes the FDO (PID) Kernel Information of an FDO (PID) record. Attributes (items) in the profile are expressed as key-value pairs where the values should be kept simple to avoid long parsing time.
FDO-KIP-P3 (from RDA-KIP-P3): The FDO (PID) Kernel Information is stored directly in the FDO record at the resolving service. If an attribute is used to enable high speed decision processes it should not be referenced to avoid additional redirection.
FDO-KIP-P4 (from RDA-KIP-P4): Change to an FDO (PID KI) record can be only by a data object owner or owner delegate (e.g., FDO (PID) record manager).
FDO-KIP-P5 (from RDA-KIP-P5): FDO (PID KI) record values should change infrequently with updates initiated only by an appropriate authority, avoiding human interaction on updates where possible
FDO-KIP-P6 (from RDA-KIP-P7): Every attribute value in an FDO profile (PID KI profile) depends only on the identified object and no other objects.
FDO-KIP-P7 (from RDA-KIP-P2): If the information for an attribute duplicates metadata maintained elsewhere, an automated synchronisation mechanism between these metadata locations has to be implemented.

Reference: Christophe, Blanchi, Maggie, Hellström, Larry, Lannom, Andreas, Pfeil, Ulrich, Schwardmann, & Peter, Wittenburg. (2023). Implementation of Attributes, Types, Profiles and Registries. https://doi.org/10.5281/zenodo.7825573

- FDO records for each digital object in DiSSCo will have the same core set of attributes, and the first one (index 1) should always be the attribute pointing to the FDO Profile, so a machine knows where to find the definition of the rest of the attributes.
- The core set of attributes is followed by other attributes. For instance, DOI Specific Elements or elements that are specific for the type of object that the pid represents. See also figure 2.
- For DiSSCo, we propose to use fixed indexes (row numbers) for the attributes, which eases the implementation of services using the information from particular attributes. The reserved ranges for the indexes can be found in table 1. For ease of use we also aim to have the attributes as a simple list of key-value pairs, without additional structure for the values. Additional information about the values like the cardinality, data type or language will be described in the FDO profile.
- The terms PID Profile, PID Kernel Profile and FDO profile are used interchangeably, the same applies to PID Record, PID Kernel and FDO record. We follow the recommendation from the FDO Forum to use the terms FDO profile and FDO record.
- The FDO record attributes purpose is for usage by machines, however some names for PIDs are included to make it also easy to understand by humans.
- Attributes use lowerCamelCase notation.

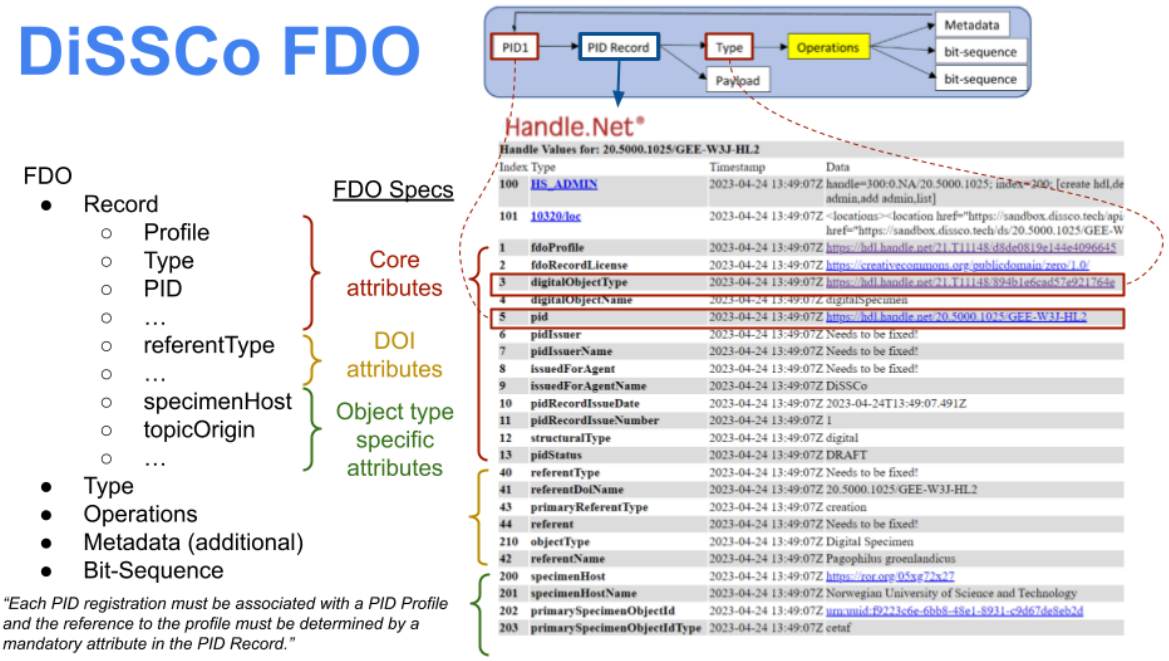

Figure 2: DiSSCo FDO

Table 1: Index guidance

1-99 series	1-29	Core FDO attributes (common within all FDOs in DiSSCo)
	30-39	Tombstones (RFC)
	40-49	DOI Specific Elements
	100-199	Handle Administration
	200-299	Digital Specimen specific elements
	300-399	Facility specific elements (facility includes collection facility)
	400-499	Media object specific elements (RFC)
	500-599	Annotation specific elements (RFC)
	600-699	Agent specific elements (source system, annotation service, user, organisation) (RFC)
	700-799	Data mapping specific elements (RFC)

	800-899	Virtual collection/dataset specific elements

Table 2: FDO Record for a Digital Specimen

Index	attribute	example	Cardinality	type
1	fdoProfile PID to a machine readable description of the attributes in the FDO record	https://hdl.handle.net/21.T11148/d8de0819e144e4096645	1/1	pid
2	fdoRecordLicense the licence for the FDO record, required to be always public domain	https://creativecommons.org/publicdomain/zero/1.0/	1/1	string
3	digitalObjectType PID to a description of the Type of digital object that defines the metadata, bit sequences (if any) and operations for the object	https://hdl.handle.net/21.T11148/894b1e6cad57e921764e	1/1	pid
4	digitalObjectName name of the object type for humans	digital specimen type 1	1/1	string
5	pid the PID of which the FDO record is part, in DiSSCo this is a Handle or DOI. It is recommended to store this pid also in the local collection management system for the specimen.	https://doi.org/10.22/GEE-W3J-HL2	1/1	pid
6	pidIssuer In case of a DOI this is a PID for the DOI Registration Agency	https://hdl.handle.net/10.17183	1/1	pid
7	pidIssuerName	DataCite	1/1	string
8	issuedForAgent In the case of a digital specimen, this is a PID for DiSSCo as the agent responsible for serving the digital specimen object	https://hdl.handle.net/10.22	1/1	pid
9	issuedForAgentName	DiSSCo	1/1	string
10	pidRecordIssueDate date the pid record was created	2022-11-24	1/1	date
11	pidRecordIssueNumber starts with 1 and is incrementally increased by 1 every time the pid record is updated. Compatible with DOI schema requirements.	2	1/1	int
12	structuralType Nature of the digital object, compatible with DOI schema requirements. The nature of a digital specimen object is always "digital". Other digital objects (outside DiSSCo) could be of physical, performance or abstraction nature.	digital	1/1	literal
13	pidStatus A PID is considered to have a lifecycle, PID status indicates the status in the life cycle, e.g. draft, active, retired. PID statuses are described further in the PID infrastructure design. One of: one of: DRAFT, ACTIVE, RETIRED, OBSOLETE, FAILED, MERGED, SPLIT	DRAFT	1/1	vocab

40	referentType A generic name for the type of object that the DOI refers to. This is different from digitalObjectType that points to a specific type, e.g. there can be different types of digital specimens that each have a slightly different metadata schema because they describe a different kind of specimen, like a botanical versus a geological specimen.	digital specimen	1/1	literal
41	referentDoiName the bare DOI Name string for the PID, e.g. without the resolver.	10.22/GEE-W3J-HL2	1/1	string
42	referentName In the case of a digital specimen this is the name for the object in the collection, which can be anything from a taxon name to a collection number.	Mus musculus type 1	1/1	string
43	primaryReferentType The primary type of the referent in the DOI Kernel XML Schema (e.g. creation, party, event). This is an open list. For digital specimens and media it will always be creation.	creation	1/1	literal
200	specimenHost ROR or, in absence of a ROR, Wikidata Qnumber for the administrative organisation hosting the specimen. Same as: Latimer Core identifierSource.	https://ror.org/0566bfb96	0/1	ROR/ Qnumber
201	specimenHostName Name for the administrative organisation as derived from ROR or Wikidata at the point of creating the FDO record.	Naturalis	0/1	string
202	primarySpecimenObjectId Primary local identifier used to identify the physical specimen, which is a preserved material entity. This identifier is usually physically attached to the specimen, e.g. as barcode. **It is recommended to use a global and resolvable identifier if available**. The host of the physical specimen should be able to find the physical specimen with only this ID plus, if that is not globally unique, the primarySpecimenObjectIdName.	RMNH.1.2b	0/1	string
203	primarySpecimenObjectIdType	Global, resolvable, local.	0/1 (mandatory if previous attribute is filled)	vocab
204	primarySpecimenObjectIdName locally used name for the identifier, to distinguish it from other locally used identifiers, max 30 characters.	registration number	0/1 (mandatory if previous attribute is filled)	string
205	normalisedSpecimenObjectId For internal processing purposes only to make a local identifier globally unique. This is the primarySpecimenObjectId if it is	0566bfb96|20.5000.1025/GEE-W3J-HL2|RMNH.1.2b		

	globally unique or a combination of ROR ID string, source system ID string, primarySpecimenObjectId if it is a proprietary identifier.			
206	specimenObjectIdAbsenceReason For example: "not yet accessioned", "digital entity specimen". Either this attribute or primarySpecimenObjectId needs to be filled, if both are filled the FDO record is invalid. Max 255 chars. Note that absence of the ID poses a challenge on avoiding duplicate digital specimen IDs.	digital entity specimen	0/1 mandatory if primarySpecimenObjectId is not filled	string

207	otherSpecimenIds
	Sequence of: id, idType, idName; with for idType the same vocabulary as primarySpecimenObjectIdType.
		{
		"otherSpecimenIds": [
		{
		"id": "12134",
		"idType": "Proprietary Identifier",
		"idName": "Specify catalog number"
		},
		{
		"id": "12134 b 34",
		"idType": "Proprietary Identifier",
		"idName": "Tissue catalog number"
		}
		]
		}
			0/many	json array

208	topicOrigin	Natural origin	0/1	vocab
209	topicDomain	Life	0/1	vocab
210	topicDiscipline	Zoology	0/1	vocab
211	topicCategory	Birds	0/1	vocab
212	livingOrPreserved	Preserved	0/1	vocab
213	baseTypeOfSpecimen	Material Sample	0/1	vocab
214	informationArtefactType (may be left blank if baseTypeOfSpecimen is not an informationArtefact)		0/1	vocab
215	materialSampleType follows the Material Sample Type vocabulary from iSamples	Whole Organism Specimen	0/1	vocab
216	materialOrDigitalEntity	Material Entity	0/1	vocab
217	markedAsType TRUE if the specimen is marked as type with a stamp or label.	FALSE	0/1	bool
218	derivedFromEntity Other material or digital entity where the specimen was derived from, specified by its ID or PID.	RMNH.1.2	0/1	id
219	catalogIdentifier		0/1	string
	follow the Material Sample Type vocabulary from iSample			
100	HS_ADMIN	handle=20.5000.1025/2FAH-GB4Y; index=300; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]	1/1	Byte array

101	10320/loc	<locations> <location id="0" href="https://sandbox.dissco.tech/api/v1/specimen/test/bff3e176-7ace-45f0-b40e-c3d8dd495de1" weight="1" view="json" /> <location id="1" href="https://sandbox.dissco.tech/ds/bff3e176-7ace-45f0-b40e-c3d8dd495de1" weight="0" view="ui" /> </locations>	1/1	Xml snippet

Note:

1. The PIDs in DiSSCo by default will redirect to the latest version the digital object itself. To access the FDO record itself use the 'noredirect' flag, for example:

http://hdl.handle.net/20.5000.1025/GEE-W3J-HL2?noredirect or
https://hdl.handle.net/api/handles/20.5000.1025/GEE-W3J-HL2?noredirect

2. Also, PIDs in DiSSCo are always Handles. DOIs are Handles with a few additional requirements and only used for objects where the PID needs to be frequently referenced by humans and long term persistence (beyond DiSSCo) is important, such as a digital specimen and a media object. Depending on the governance model and RA service we adopt, DOI minting can have additional costs and limitations in performance.

3. DOI also have support for describing the IDs:

- returnType, e.g. "text/html" or "application/json"
- doesContentNegotiation
  We could include these for the specimen identifiers but may not be useful enough for machines to be worth the effort to try to get that data in.

Benefit

- Separation PID level metadata vs content level metadata
- Large scale machine actionability based on a minimum set of attributes that can be common across different digital objects
- Abstraction: every FDO within DiSSCo can be treated the same at a certain level until needed to be treated differently

Competition/Alternatives

- We could implement the RDA recommendation on PID Kernel Information. The proposed list of core attributes in this RFC was based on RDA recommendation and also on DOI and DataCite schema elements. The RDA recommendation was not exactly implemented as it has certain shortcomings compared to the needs of

DiSSCo. For example, our digital objects are mutable, so relations with other objects may be added and change over time and would fit better in the metadata part of the object itself. Also, the objects can be containers including other digital objects (e.g. the digital specimen including image objects, or a virtual collection set of digital specimens).. These may have different licences, which makes it hard to provide licence information for the object itself in the FDO record and it may also change if new objects are added to the container object. On the other hand we found it important to add a licence for the PID record itself, even though that will always be public domain, just to make that explicit for future use as open data. Also, we may want to include separate checksums e.g. content and metadata parts.
To make it concrete, these attributes will be part of the digital object metadata instead:

- digitalObjectPolicy
- etag (Checksum of object contents).
- version (we always serve the latest version)

And relationships with other PIDs will also be part of the metadata as they will change over time. Here we will not follow the kernel recommendation which included:

- wasDerivedFrom
- specializationOf
- wasRevisionOf
- hadPrimarySource
- wasQuotedFrom
- alternateOf

- We could rely (only) on JSON schema enforcement or RDF description of the data. That would make the data machine readable but not actionable: possible operations would not be described. It is currently unlikely that machines would be able to discover and use APIs without guidance from developers where to find them and how to use them.
- We could use the PID record only for information about the location to redirect to, as currently common for DOIs. That would require putting all metadata together into a separate location or include it in the object (e.g. a digital specimen).The distinction between content level metadata and data is difficult to make. Also, then a machine would need to retrieve the full object first before it can assess whether it can use or further act upon it.

Limitations

- We assume that there is always one administrative organisation responsible for hosting the specimen. If there is more than one, then we should find a way to identify the primary one and include that value in the record.
- The separation character used in “normalisedSpecimenObjectId" cannot be used in the proprietary identifier. Additionally, in cases where there are multiple proprietary identifiers within a source system, we don't know which one from the source system was used and can only find out through the mapping file of the source system (or the name of the identifier). We also assume that the source system does not change

often, and that for new specimens collected or new digitisation projects a globally unique identifier will be used. Source system PID could be left out if the proprietary identifier is unique within the hosting organization, but we do not know for sure other than from information provided by the organization, so it may be better to always include it.

- There is no global index for FDO records. However, if a digital object is part of an index like DiSSCo, a machine can use that index to find other objects or related objects that are included in the index, through operations described in the type definition.
- Governance and maintenance of the FDO record can be extra overhead and there are no machines yet that make use of it.

Appendix: Vocabularies

Idx	Term	Accepted Terms
12	structuralType	digital, physical, abstraction, performance
13	pidStatus	Draft, Active, Retired, Failed, Merged, Split, Obsolete
40	referentType	Digital specimen, Agent, Digital Media, Facility, Annotation, Virtual collection, Dataset mapping
43	primaryReferent Type	creation
203	primarySpecimenObject IdType	Global, Local, Resolvable
208	topicOrigin	Natural, Human-made, Mixed origin
209	topicDomain	Life, Environment, Earth System , Extraterrestrial, Cultural Artefacts, Archive Material
210	topicDiscipline	Anthropology, Botany, Astrogeology, Geology, Microbiology, Palaeontology, Zoology, Ecology, Other Biodiversity, Other Geodiversity
211	topicCategory	See Specimen and Collection Classification
212	livingOrPreserved	Living, Preserved
213	baseTypeOfSpecimen	Material entity, Information artefact
214	informationArtefactType	DCMI:Image, DCMI:MovingImage, DCMI:PhysicalObject, DCMI:Sound, DCMI:3DResourceType
215	materialSampleType	See Specimen and Collection Classification
217	markedAsType	true, false