Enabling Discovery (CPP-024)

CPP-Identifier	CPP-024
CPP-Label	Enabling Discovery
Author	Bertrand Caron
Contributors	Mattias Levlin
Evaluators	Matthew Addis, Felix Burger, Maria Benauer
Date of edition completed	29.08.2025
Change history	**Comments**
Version 1.0 - 29.08.2025	Milestone version

1. Description of the CPP

The TDA provides catalogue services to its consumers to help them identify *Objects* that they may be interested in.

Inputs and outputs

Input(s)	
Data	Optional: Derivatives useful for the consumer to identify the content and scope of the *Object*
Metadata	Digital Archive Database
Documentation / guidance	Query scenarios
	Metadata mapping specifications
Output(s)	
Metadata	Catalogue service
	Process execution report

Definition and scope

Enabling Discovery covers the extraction of the subset of *Metadata* from the digital archive database to enable consumers to identify which *Objects* they may be interested in. In OAIS, this subset is called "Descriptive Information".

Discovery relies on CPP-018 (**Community Watch**) to study consumers' needs regarding searches and elaborate query scenarios. Query scenarios are use cases where the consumers identify a subset of the TDA holdings' that addresses their needs. Discovery is in charge of providing *Metadata*, query and retrieval features that support these query scenarios in an efficient way. If the TDA wants its *Objects* to be discoverable in third-party catalogues (e.g. federated catalogues, portals etc.), it must perform this process for each service in order to ensure that *Metadata* and derivatives conform to the specifications of the third-party catalogue(s).

If the TDA provides direct access to its holdings for end users, then it may also generate a *PID* (e.g. a DOI) for each accessible *Object* so that it is referencable and locatable using third-party discovery services. However, this CPP only focuses on the generation of *Metadata* that is subsequently used for discovery (rather than the discovery process and the use of discovery services such as portals and catalogues). The details of registering PIDs, publishing *Metadata* and/or data with third-party catalogues and discovery services, as well as the access to *Objects* via PID resolution and TDA services are beyond the scope of this CPP.

Enabling Discovery is also in charge of verifying the legal status of *Metadata* based on rights assessment as issued by CPP-020 (**Rights Management**). *Metadata* access might indeed

need to be restricted, so Discovery must ensure it disseminates these only to authorised users.

In addition to *Metadata*, discovery might provide derivative copies (e.g. thumbnails, textual transcription, redacted copy, etc.) that would help consumers in identifying the content and scope of the relevant *Objects*. Indeed, access to the preservation copies or to a sufficiently complete derivative is often limited to the TDA's precinct.

Process description

Trigger event(s)

Trigger event	CPP-identifier
New *Object* or new *Object* version ingested	CPP-029 (Ingest), CPP-021 (AIP Versioning)

Step-by-step description

No	Supplier	Input	Steps	Output	Customer
1	CPP-018 (Community Watch)	Query scenarios	Select the relevant *Metadata* (and possibly data) useful to the consumer	Subset of *Metadata* to be exposed in the catalogue	
2	CPP-018 (Community Watch)	Query scenarios	Select the syntax or serialisations useful to the consumer	Syntax of *Metadata* to be exposed in the catalogue	
3a		New *AIP* or *AIP* version	Extraction: Extract the required subset of *Metadata* from the new *AIP* or *AIP* version	Extracted discovery *Metadata*	
		Subset of *Metadata* to be exposed in the catalogue			
		Syntax of *Metadata* to be exposed in the catalogue			
3b		Extracted discovery *Metadata*	Mapping & transformation: Map and transform the *Metadata*	Transformed discovery *Metadata*	

			according to the required format of the discovery catalogue		
3c		Transformed discovery *Metadata*	Validation: Validate the resulting subset of *Metadata* against the target catalogue's schema	Validation successful: Validated discovery *Metadata* (step 4)	
				Validation failed: Log the error and flag the record for review. The record must not proceed to the discovery catalogue until corrected	
4	CPP-020 (Rights Management)	Validated discovery *Metadata*	Check rights status of discovery *Metadata*	Validated discovery *Metadata* with cleared rights	
		Rights statement			
5a		Validated discovery *Metadata* with cleared rights	Add discovery *Metadata* for the *Object(s)* to the catalogue service. The catalogue service may be provided by the TDA, by a third-party (e.g. a federated catalogue), or by a combination of the two.	Entry in catalogue service	Consumer
5b		Validated discovery *Metadata* with cleared rights	Optional: If the TDA requires a PID (e.g. DOI), because the *Object* or *Metadata* about the *Object* will be publicly accessible (e.g. open access), the TDA may request a PID from an appropriate	PID	Consumer

			registration agency. The TDA may also add the PID to the *Metadata* in the catalogue service.		
5c	CPP-025 (Enabling Access), CPP-028 (Creation of Derivatives)	*DIP* or Derivatives	Optional: The TDA may provide versions of its *Objects* for inclusion in the catalogue service (for example, thumbnails, preview versions, redacted documents etc.). These are added to the catalogue entry for the *Object*.	Entry in catalogue service.	
6		Query scenarios	Verification: Using the defined query scenarios, perform a test query to confirm that the new or updated *Object* is discoverable in the catalogue service.	Verification successful: Verification confirmation in the process execution report	CPP-013 (Object Management Reporting)
		Catalogue service		Verification failed: If the *Object* is not found, log the failure as an incident to investigate the ingest and indexing chain	

Rationale(s)[1] and worst case(s)

Rationale	Impact of inaction or failure of the process
If the TDA is providing access to its holdings to consumers, ensuring query services that suit the consumers' needs is mandatory.	If the catalogue service does not allow query scenarios useful to the consumers, discoverability and usage of the TDA holdings is compromised.

2. Dependencies and relationships with other CPPs

Dependencies

CPP-ID	CPP-Title	Relationship description
CPP-005	Identifier Management	Enabling Discovery should make use of PIDs.
CPP-009	Metadata Extraction	Some Metadata provided to the consumer must have been extracted from the Files.
CPP-016	Metadata Ingest and Management	Enabling Discovery relies on a correct metadata management process. In particular, *Metadata* created by and within the TDA is of particular interest to the consumer in order to understand preservation actions that could have affected the *Object*.
CPP-018	Community watch	The TDA must have identified the needs of its designated community in order to enable queries that support the community's defined query scenarios.

Other relations

Relation	CPP-ID	CPP-Title	Relationship description
Affinity with	CPP-025	Enabling Access	The distinction between **Enabling Discovery** and **Enabling Access** may be blurred as derivative copies may be indexed and searched in the same way as *Metadata*. In addition, these derivative copies may be sufficient to address some

---

[1] Term derived from PREMIS.

			consumers' needs. Nevertheless, the distinction is still useful as giving access to the original data is often governed by specific legal constraints, and requires specific hardware and software tools.

3. Links to frameworks

Certification

Certification framework	Term used in framework to refer to the CPP	Section
CTS Link	Discovery	R12 Discovery and Identification
Nestor Seal Link	Research (options)	C4 Access
ISO 16363 Link	Discovery	4.5.1 The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest.

Other frameworks and reference documents

Reference Document	Term used in framework to refer to the process	Section
OAIS Link	*No exact term is available in OAIS for Discovery, but the topic is approached through the* Package Description *notion*	4.3.3.7 Unit Description 4.3.3.9 Collection Descriptions
PREMIS Link	Discovery	Section "More on Objects", subsection "Intellectual Entities", p. 8.

4. Reference implementations

Example use cases

ePADD Discovery module

Institutional Background	
Institution	Stanford University's Special Collections & University Archives, USA
Hyperlink	https://www.epaddproject.org/using-epadd/discovery-module Discovery module for Stanford's email collections: https://epadd-discovery.stanford.edu/epadd/collections
Description	
Problem statement	Email collections are born-digital material that need and allow specific usages and access. On the other hand, they raise privacy issues that force memory organisations to give access to these collections only in the organisation's precinct. Consumers therefore need to identify the scope and content of the collection through remote queries before planning an on-site visit.
Proposed solution	Beyond the mails' *Metadata*, the full text was indexed and is searchable, though not entirely readable - when accessing the mail, the full text is redacted, only the searched term and email *Metadata* are displayed. Access to the mails' collections is possible only in the organisation's precinct.

etsin.fairdata.fi (CSC): Open discovery for datasets with varied access conditions

Institutional Background	
Institution	CSC – IT Center for Science, Finland
Hyperlink	Main Discovery Service: https://etsin.fairdata.fi/ Example 1 (Direct Access via use-copy): FIRE Profile 3 Example 2 (Mediated Access): Drive-tested topsoil
Description	

Problem statement	A TDA can host datasets with a wide variety of access conditions. Some may be open for immediate download (e.g. via a use copy), while others may require human-mediated access (even if their ultimate license is open). The challenge for a discovery service is to represent a) these different states, b) providing access pathways for each dataset without compromising the specific rights management and c) providing access policies defined by the data owner. The Finnish discovery portal etsin.fairdata.fi (CSC) hosts *Metadata* for both active research data outside of the AIP context and for formally preserved assets (*AIPs*). The first challenge is to provide a clear and reliable method for users to discover *only* the preserved assets (with associated AIPs); and second, to represent the specific access conditions for each of those preserved *AIPs*.
Proposed solution	The Fairdata services solve this by addressing the discovery of preserved assets by providing a specific, filterable view of its catalogue. The URL parameter ?data_catalog__title=Fairdata+PAS+datasets in https://etsin.fairdata.fi/ creates a dedicated view within Etsin that exclusively lists datasets corresponding to verified AIPs in the long-term Digital Preservation Service (the filter is also selectable via the Etsin user interface). Within this view where only datasets with an associated AIP are visible, the system handles varied access policies, as defined by the data owners. It can enforce a spectrum of access conditions: Example 1 (Direct Access): For some *AIPs*, a "use copy" is made available for immediate download. Etsin provides a direct link to this copy while also linking to the canonical AIP record (which is not available via Etsin), separating simple use from the formal preservation record. Example 2 (Mediated Access): For other *AIPs*, access is controlled. Etsin displays a record that might offer a link to a "preview dataset" but withholds the full data. It then provides explicit instructions on the required procedure, such as "To get this dataset, please contact [email address]." Example 3 (Fully Mediated Access): The system can also handle cases where no preview or data is available, and the only option for a user is to follow the human-mediated contact instructions provided in the *Metadata*.

| | These examples demonstrate how a discovery service can provide a dedicated view for preserved *AIPs* (CPP-024) while interpreting and enforcing a range of different access rules—from fully open to fully mediated—as defined by rights and access policies (CPP-020) and taken into account during the access process (CPP-025). |

Publicly available documentation

Institution	Organisation type	Language	Hyperlink
TIB – Leibniz Information Centre for Science and Technology and University Library, Germany	National library	English	Internal use: https://wiki.tib.eu/confluence/spaces/lza/pages/93608951/Metadata, also see https://knowledge.exlibrisgroup.com/Rosetta/Training/Rosetta_Essentials/Data_Management/7.1_Searching_the_Rosetta_Permanent_Repository
	Non-commercial digital preservation service		External Use: TIB Portal: https://www.tib.eu/en/, TIB AV Portal: https://av.tib.eu/, Library union catalogue k10plus: https://gvk.k10plus.de/LNG=EN
	Research infrastructure		
	Research performing organisation		