

---

# FINDING AND USING DIGITALLY PRESERVED ASSETS: A CONSUMER-CENTRIC PERSPECTIVE

---

Kenneth Thibodeau

July 2010

---

## ABSTRACT

---

This document explores what is necessary and beneficial for a consumer to discover and use digitally preserved assets, identifies challenges to discovery and use, and suggests ways the digital preservation interoperability could contribute to meeting individual challenges. It assumes that a consumer's needs for information and for processing information are determined by the consumer's intended use; and that most often this will be independent of whether the information is preserved or current. It also assumes that consumers generally will want to use current standards, tools and processes to discover and use digitally preserved assets.

---

## DEFINITIONS OF TERMS

---

**Collection:** A set of Information Objects assembled by a person, organization or service to serve a specified purpose.

**Consumer:** The role played by those persons, or client systems, who interact with digital preservation services to find preserved information of interest and to access that information in detail. Derived from CCSDS, 2002

**Descriptive Information:** (OAIS) "The set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding, ordering, and retrieving of OAIS information holdings by Consumers." (CCSDS, 2002)

**Digitally Preserved Asset (DPA):** a digital information object that has been designated for long-term preservation

**Digital Preservation Domain (DPD):** the region of cyberspace where digital information is preserved and accessible and the superposition over this space of structures and rules that manage preserved information and processes that act on or with it.

**Digital Preservation Interoperability Framework (DPIF):** a comprehensive and coherent set of standards, specifications and guides for communication, including invoking services and returning

results, across the external External of the DPD and between and among actors and entities within the DPD.

**Dissemination Information Package (DIP):** “The Information Package, derived from one or more AIPs, received by the Consumer in response to a request.” Derived from CCSDS, 2002

**Information Object:** “A Data Object together with its Representation Information.” (CCSDS, 2002)

**Preservation Description Information (PDI):** “The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, and Context information.” (CCSDS, 2002)

**Producer:** “The role played by those persons, or client systems, who provide the information to be preserved.” (CCSDS, 2002)

**Repository:** a service which maintains information assets and provides access to them. (Wilensky, 2001) (Kahn, 1995)

**Representation Information:** : The information that maps a Data Object into more meaningful concepts.” (CCSDS, 2002)

## INTEROPERABILITY CHALLENGES IN CONSUMER REQUIREMENTS

Information about the challenges consumers face in finding and using digitally preserved information and ways that interoperability could contribute to meeting those challenges is presented in table form. While consumer requirements for discovery and use are related to repository and even producer requirements, only consumer requirements are considered here.

The table is organized in the following columns: Consumer Requirement, Challenge, How Interoperability Could Help, and the Locus of Interoperability. The contents of each column are described below.

### CONSUMER REQUIREMENTS

- Discovery: the consumer can learn of the existence of DPAs and easily determine a path to them, if desired.
- Relevance: the consumer can determine whether and how the information could be used to accomplish the stated purpose.
  - Content
  - Data Type: metadata about data type (e.g., quantitative v. qualitative, alphanumeric v. image, structured v. semistructured v. unstructured) should is available
    - The consumer needs to select a subset of the most promising information.

- Appropriateness: an overview indicating salient aspects of the information, including summary indications of adequacy, precision, reliability and usefulness, to enable an initial assessment of the probability that the information will contribute to the intended purpose.
- Adequacy: there is enough information to make it worthwhile.
- Precision: the information is at an appropriate level of detail
- Reliability
  - Accuracy: the information is verifiably correct or aspects that are erroneous, misleading or not verifiable are specified and tolerable
  - Trustworthiness of the producer: the producer is believed to be competent
  - Sufficiency and appropriateness of the data provenance: enough is known about how the data has been generated and modified to support a conclusion that it can be used for the intended purpose.
  - Trustworthiness of the preservation: the chain of preservation is known and provides sufficient basis for believing the information has not been corrupted
  - Sufficiency and trustworthiness of the representation information and contextual information to enable correct interpretation of the data
- Usefulness
  - Ease of acquiring or accessing the information
  - Suitability of the format for the intended use
  - Ease of performing any needed processing
  - Possibility of combining the information with other information.

There are many ways in which interoperability is implicated in determining whether these criteria are satisfied.

---

### CHALLENGE

Each entry describes some difficulty in satisfying a requirement in the first column

---

### HOW INTEROPERABILITY COULD HELP

Each entry describes how an aspect of interoperability could help in meeting a particular challenge.

---

### LOCUS OF INTEROPERABILITY

This column identifies where interoperation would need to occur. The domain includes:

- Cyberspace: generic across cyberspace, including but not limited to the DPD
- DPD: generic across the DPD
- External: at or across the external between the DPD and the rest of cyberspace
- X: DPD entity of type, X
- $X \rightarrow Y$ : Between or among entities of type, X, and entities of type, Y

## REFERENCES

---

CCSDS, C. C. (2002). *Reference Model for an Open Archival Information System (OAIS)*. Retrieved from <http://public.ccsds.org/publications/archive/650x0b1.pdf>

Kahn, R. and Wilensky, R. (2006). A Framework for Distributed Digital Object Services. *International Journal of Digital Libraries*, 6 (2), 115-123.

Wilensky, R. (2001). *Personal libraries: Collection management as a tool for lightweight personal and group document management*. . Technical Report to the National Archives and Records Administration, San Diego Supercomputer Center.

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability
Discovery	Consumer purpose may be served by current as well as preserved assets	DPAs should be discoverable using generalized pathways and tools that operate across cyberspace	External
		DPIF should define an interface that facilitates external search tools operating against a variety of interfaces in preservation repository	
	Discovery of DPAs may be hindered by changes in terms over time	Discovery tools could suggest terms related to those entered by consumers	Cyberspace
		DPIF should define implement a standard syntax for expressing precedence – succession relationships, including cardinality and chronology, for faceted search	DPD
	Assets identified in search may include objects which either are or appear to be copies or versions of one another	DPIF should provide a standard for associating copies, versions, and derived products	DPD
		Repositories should implement DPIF standard for associating copies, versions and derived products located within their collections and in other repositories	Repository ↔ Repository
		Consumers should implement DPIF standard to associate copies, versions and derived products with source DPAs, submitting or linking this information to the repositories that provided the DPAs	Consumer → Repository
	Copies or versions of a DPA may be located both within and outside of the DPD	DPIF standard for associating copies, versions and derived products should be consistent with or capable of two way translation to other methods used for this purpose	External
	Relevant DPAs may be located in several repositories	DPIF should provide basic syntax and terminology for queries and orders for DIPs, allowing specializations appropriate to the types of information and the Designated Communities of each repository	Repository ↔ Consumer

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability	
	Result set may include information assets derived from DPAs such that some significant portion(s) of the processing needed for the consumer purpose was accomplished in production of the derived asset(s)	The DPIF standard for associating variants of a DPA should readily link to data provenance information for each variant.	Repository ↔ Consumer	
	Consumer may not know where the information is held	Possibility of discovering DPAs should not depend on the repository. A standard method to link from information identifying or characterizing a DPA to a repository which can provide access to it should be part of the DPIF	DPD	
Relevance	For many purposes, information about assets returned by Internet search engines is not sufficient to determine relevance	DPIF should endorse the Descriptive Information model of OAIS and provide guidance on appropriate levels and types of Descriptive Information, beyond the minimum allowed by OAIS, to enable consumers to determine relevance for different purposes.	DPD	
	Norms for descriptive information will vary according to several factors; such as, discipline, type of institution (scientific data center, government archives, university library, health care provider), designated community, et al.	DPIF could establish a standard for common descriptive information to be provided about all DPAs.	External	
		DPIF could provide a service to translate variant descriptive information to a coherent form, without hiding important idiosyncrasies.	External	
	Standards for descriptive information will change over time.	DPIF should provide a mapping between data elements in descriptive standards.	DPD	
Content	Result list DPAs may be in multiple repositories	DPIF should facilitate requesting copies of, or access to, DPAs that appear on a results list but are preserved in different repositories	DPD → Repository	
	Data Type	Inconsistent or inadequate characterization of data type	DPIF should implement a taxonomy of data types	DPD
		Repositories should implement DPIF data type taxonomy	DPD	Repository
		DPIF data type taxonomy should be consistent with similar structures used outside the DPD	External	

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability
Appropriateness	Consumer needs to make an initial assessment of whether the DPA will serve the intended purpose	DPIF should endorse the Preservation Description Information (PDI) model of OAIS and require that sufficient PDI be accessible, independently of a DIP, to enable consumers to assess appropriateness of DPAs	DPD
Adequacy	Consumer purpose may require aggregating data from different sources to provide consistent coverage over a length of time.	Utility to display chronological coverage of an arbitrary set of assets on a time line	DPD
		Ability to create, maintain, and modify a virtual collection, drawn from several repositories, organized according to temporal coverage. Possibility for other users to discover and use such virtual collections.	Cyberspace
	Possibility of using a DPA of a given data type may depend on availability of information in another data type; e.g., encoded data requires codebooks for interpretation; textual documents may require external font libraries; etc.	Standards for preserving relationships among DPAs	DPA set
Precision	Syntax and semantics for specifying precision may vary	Repositories should implement standard methods for specifying precision appropriate to a given discipline or content domain	Repository
		DPIF should include an abstract grammar for specifying precision suitable for tracking changes in lower level specifications in given domains over time.	DPD
Reliability	Reliability may vary substantially across a set of information assets derived from multiple sources	DPIF should provide a common syntax and semantics for repositories to communicate with producers and consumers about reliability	Producer ↔ Repository ↔ Consumer
	Different purposes may entail different ways of determining reliability	DPIF should facilitate the use of standard, discipline-specific approaches to determining reliability for different purposes and different data types	DPD
		DPIF should provide or utilize a meta-semantics for characterizing different approaches to determining reliability.	

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability
Accuracy	DPAs with related content may have different levels of accuracy		
Producer	Promising DPAs may come from several producers	DPIF could provide standards for metadata about producers	Repository
Data Provenance	The chain of preservation or other preservation data may not adequately or accurately identify producers or processes that brought the data to its preserved state.	DPIF should include minimum standards for data provenance for different data types Beyond the minimum, DPIF should facilitate preservation and communication of data provenance metadata as specified in standards applicable to different disciplines	DPIF
Preservation	Information could be corrupted over time.	DPIF should provide standards for describing how DPAs have been preserved	DPA
		MOIMS-RAC supplement to OAIS standard will provide basis for assessing the trustworthiness of a repository	Repository
	Information about RAC certification of trustworthy repositories may not be readily available		Repository
	There are likely to be variations in repository certification information both between different repositories and over time.		External
	Certification of trustworthiness of a repository may not be sufficient to determine that a DPA in that repository has been appropriately preserved.	Repositories should provide preservation information about DPAs	Repository → Consumer
Interpretation	Various factors, such as differences in the ways data was created or collected, variations in the performance of instruments, et al., can be important for valid interpretation of data	Repositories should identify the types of factors which can influence interpretation of data within the scope of their collections and motivate producers to report accordingly.	Repository → Producer
		DPIF should define standard tags for indicating the presence of idiosyncratic factors impacting on interpretation.	DPIF

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability
		DPIF should provide guidance on how to indicate interpretive considerations in PDI and Representation Information.	
Usefulness			
Acquiring or accessing	Are there restrictions on release, use, or downstream dissemination, including derived products	Standard formulation of restrictions that is invariant regardless of producer or repository	DPD
Acquiring	Does the volume of an asset impede physical transfer?	Repositories should facilitate access to and processing of data within the repository when data volume or other factors make transfer difficult	
	Does the volume of a set of assets impede physical transfer?		
	Packaging of DPAs in AIPs or DIPs may be ill suited to a consumer's needs	DPIF should provide guidance on approaches to packaging DPAs that facilitate discretionary selections and joins	Repository
Accessing	Can an asset be accessed and desired processing performed within the repository?	DPIF should provide standard syntax and semantics for identifying constraints.	
	DPAs that cannot easily be transferred may be in different repositories with different possibilities for access and in situ processing		
Format	Does the repository have a coherent approach for identifying formats from different generations of technology?	DPIF should have a common and comprehensive taxonomy of data formats	DPD
		DPIF data format taxonomy should include current as well as obsolete formats. Entries for current formats should be unambiguously linkable to corresponding metadata used to identify and characterize the same formats outside of the DPD	Cyberspace

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability	
	Does the repository have a coherent approach for identifying formats when the relevant assets came from multiple producers?	DPIF data format taxonomy should provide for local extensions	DPD	
		DPIF taxonomy should have a service for identifying duplicate or similar format captured as local extensions	DPD → repositories	
		DPIF taxonomy should facilitate elevation of duplicate local extension to DPD level		
		Different repositories may use different ways to identify and characterize digital formats.	DPIF should define or adopt a metadata standard for exchange and dissemination of information about digital formats	DPD
		Various disciplines and communities of consumers may have different needs for information about formats		
	Processing	Can tools for performing desired processing be applied to the target DPAs?	DPIF ontology should provide for specifying the domain of format(s) that can be input or output of a tool.	DPD
	DPIF Ontology should include a taxonomy of processing services including transformation, data interlinking, integration, fusion, analysis, annotation, presentation, and others			
	DPIF taxonomy of processing services should identify which services can be applied to common formats of DPAs without transformation.			
	The DPIF taxonomy of services should be consistent with related taxonomies in wide use in cyberspace		Cyberspace	
	Consumer requirements may entail use of distributed capabilities such as computational grids, data grids, et al.		The DPIF should include standard service interfaces that enable use of distributed external resources for processing of DPAs	External
			DPIF should specify standard interfaces for OAIS compliant repositories to connect to grid resources	Repository
	Processing of sets of DPAs may require special capabilities for temporal interlinking, integration or fusion		DPIF should specify standard translation of different chronological and temporal scales	DPD
			DPIF should define common services for temporal	

Consumer Requirement	Challenge	How Interoperability Could Help	Locus of Interoperability		
		integration or fusion			
		DPIF should define standard service interfaces for mediation services for resolving semantic and syntactic inconsistencies specifically related to the temporal dimension			
		DPIF should specify interface for invoking semantic and syntactic mediation services available in cyberspace	External		
		If selected processing cannot be performed on DPAs in the formats in which they are preserved, can the DPAs easily be transformed to suitable formats?	DPIF ontology should include identification of feasible transformations with characterization of data loss or other changes entailed by each transformations	DPD	
		Tools may not be able to process all the formats in which similar content exists	DPIF should facilitate the creation, continuing enrichment, and sharing of registries of metadata about the suitability of tools useful for processing DPAs	DPD	
		If the assets are located in multiple repositories and some must be processed in their current locations, can desired processing be performed consistently on all target assets?	DPIF	Consumer → Repository	
		Combination	Consumer may need to create data integration, fusion, or other combinations of information of a variety of both preserved and current information assets	DPIF should facilitate discovery, by repositories or consumers, and determination of appropriateness of services for various combinations of data including multiple instances of data of a single data type or format and of heterogeneous data types and formats.	Cyberspace
		DPIF should facilitate sharing of information about data combination services and their performances, especially with respect to data types and formats that are more frequent or important in the DPD than in cyberspace generally		DPD	