**The case for a revision of OAIS**

From <http://wiki.dpconline.org/index.php?title=The_case_for_a_revision_of_OAIS> by [David Rosenthal](http://wiki.dpconline.org/index.php?title=User:DRosenthal)

**COMMENTS by David Giaretta**



[**Community Forum**](http://wiki.dpconline.org/index.php?title=A_community_forum_for_digital_preservation_and_curation_standards) **|** [**OAIS Community**](http://wiki.dpconline.org/index.php?title=OAIS_Community) **|** [**OAIS Structure**](http://wiki.dpconline.org/index.php?title=OAIS_Structure) **|** [**OAIS Blog Posts**](http://wiki.dpconline.org/index.php?title=OAIS_Blog_Posts) **|** [**Active Topics and News**](http://wiki.dpconline.org/index.php?title=Active_Topics_and_News)

The official title of ISO 14721 is [*Reference Model for an Open Archival Information System (OAIS)*](http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284). The role of a reference model is to provide abstract concepts and terminology by means of which concrete systems can described and analysed. A reference model is not of itself a standard against which concrete systems can be assessed for conformance, that is the role of criteria based on these concepts and terminology. In the case of ISO 14721 this role is performed by [ISO 16363](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510) and its predecessor TRAC. The effectiveness of ISO 14721 must be judged by the effectiveness of its concepts and terminology in describing concrete archival systems, and audits under TRAC and ISO 16363 provide a valuable opportunity to do so.

In July 2014 the [CLOCKSS Archive](http://www.clockss.org) was [certified by CRL](http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/clockss-report) after a rigorous audit against the TRAC criteria, the process for certification under ISO 16363 not then being available. CLOCKSS gained an overall score that equalled the previous best, and the first ever perfect score in the "Technologies, Technical Infrastructure, Security" category. All non-confidential documents submitted to the auditors are available [here](http://documents.clockss.org). Four blog posts describe [the certification](http://blog.dshr.org/2014/07/trac-certification-of-clockss-archive.html), [the audit process](http://blog.dshr.org/2014/08/trac-audit-process.html), [the lessons learned](http://blog.dshr.org/2014/08/trac-audit-lessons.html), and [how to run the demonstrations](http://blog.dshr.org/2014/08/trac-audit-do-it-yourself-demos.html) we showed the auditors.

In general, basing the description of the CLOCKSS Archive on the ISO 16363 criteria, and thus on the concepts and terminology of ISO 14721 worked well. Documents describing in detail the way significant OAIS concepts apply to the CLOCKSS Archive are available [here](http://documents.clockss.org/index.php/Main_Page#OAIS_Conformance_Documents). But the ["lessons learned"](http://blog.dshr.org/2014/08/trac-audit-lessons.html) blog post includes a section OAIS vs. CLOCKSS, reproduced here:

Writing the OAIS Conformance Documents made the mis-match between the theory of the OAIS reference model and the practice of digital preservation in the Web era, and in particular that of the CLOCKSS Archive, evident. The conceptual mis-matches between the OAIS Reference Architecture, upon which ISO 16363 is firmly based, and the CLOCKSS Archive's architecture fall into four broad areas:

* **CLOCKSS is a dark archive**. Eventual readers of the archive's content are unknown, and have no influence over when, whether and how content is released from the archive. The OAIS concept of Designated Community is thus difficult to apply.
	+ - **COMMENT: This is a misunderstanding of the definition of Designated Community. The Designated Community is defined by the archive. The archive does not have to see into the future – they just have to make it clear what they are doing. The “eventual users” may or may not be part of that Designated Community, and are not required to have any influence on when, whether and how content is released. The archive will have some process for making these decisions but OAIS does not cover those.**

* **CLOCKSS ingests streams of content**. Content ingested by crawling the Web, as much of the CLOCKSS Archive's content is, is not pushed from the content submitter to the archive but pulled by the archive from the publisher. The publishers of academic journals emit a continual stream of content; any division into units is imposed by the archive, not by the publisher. The OAIS concept of Submission Information Package, (SIP) and the relationship it envisages between the submitter and the archive, is difficult to apply. The concept of Archival Information Package (AIP) also has some detailed mis-matches, since to collect a stream an AIP must be created before it contains any content, and subsequently accumulate content over time instead of, as OAIS envisages, being wrapped around a pre-existing collection of content at creation time.
	+ - **COMMENT: The AIP is certainly defined by the archive. The SIP is a general concept and the Producer is a role rather than a specific person or organisation. The AIP is not assumed to be created before there is any content – that would be most peculiar. The archive decides how to create the AIP. OAIS specifies the kinds of information which must be logically contained in it.**
* **CLOCKSS has a centralized organization but a distributed implementation**. Efforts are under way to reconcile the completely centralized OAIS model with the [reality of distributed digital preservation](http://purl.pt/24107/1/iPres2013_PDF/Creating%20a%20Framework%20for%20Applying%20OAIS%20to%20Distributed%20Digital%20Preservation.pdf), as for example in collaborations such as the [MetaArchive](http://www.metaarchive.org/) and between the [Royal and University Library in Copenhagen](http://www.kb.dk/en/) and the [library of the University of Aarhus](http://library.au.dk/en/). Although the organization of the CLOCKSS Archive is centralized, serious digital archives like CLOCKSS require a distributed implementation, if only to achieve geographic redundancy. The OAIS model fails to deal with distribution even at the implementation level, let alone at the organizational level.
	+ - **COMMENT: The Functional Model is a logical representation, not a design for a centralised archive. OAIS does not specify how the various Functional Entities are implemented or distributed. Standards for various aspects of implementations would be better placed in a separate standard which follows the OAIS Reference Model concepts and terminology.**
* **The CLOCKSS Archive contracts-out its operations**. The CLOCKSS Archive not-for-profit achieves its low cost of operations by contracting them all out under two contracts with Stanford University. This enables many costs to be shared with the other users of the LOCKSS technology, to the benefit of both. The OAIS model fails to deal with organizational divisions such as this.
	+ - **COMMENT: Again the Functional Model does not specify how the Functional Entities are implemented.**

Another mis-match between OAIS and web archiving would have been a problem had CLOCKSS not been a dark archive. Access to archived Web content, via [Memento (RFC7089)](http://www.mementoweb.org/), direct link or text search, occurs at the level of an individual URL. The OAIS concept of Dissemination Information Package is difficult to apply to access of this kind; it says:

In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a Dissemination Information Package (DIP). The DIP may also include collections of AIPs, and it may or may not have complete PDI. The Packaging Information will necessarily be present in some form so that the Consumer can clearly distinguish the information that was requested. Depending on the dissemination media and Consumer requirements, the Packaging Information may take various forms.

Although there is obviously a lot of room for interpretation here, it does not appear to cover the case where the Consumer requests, and the archive delivers, a digital object (the headers and body of a URL) in exactly the form it was ingested with no Packaging Information. This is what Consumers of archived Web content want. It is true that, for example, Memento adds header information to its response, but that information serves to point to other archived digital objects, potentially in other archives, so it can't be considered Packaging Information for the requested DIP. Fortunately for us, the trigger process of the CLOCKSS Archive does deliver a package containing many URLs, so it more closely matches the OAIS DIP concept.

**COMMENT: The DIP is again a general concept and OAIS does not say how any particular DIP is constructed or what it will contain. However, it must be possible for an archive to provide the details of how the information in the DIP links back to the original information which the archive ingested. However not all DIPs need to contain that provenance. Packaging Information is defined as:** *The information that is used to bind and identify the components of an Information Package*. **If the response (the DIP) is sent using HTTP then the information that it is HTTP is the Packaging Information – normally taken care of by the browser without the knowledge or intervention of the human user.**

Our experience in the TRAC audit of the CLOCKSS Archive reveals a number of areas in which the concepts and terminology of ISO 14721 are inadequate to describe a real, functioning system. There are two ways to react to this. If you believe that ISO 14721 is not a *reference model*, but a *definition* of an archival system, your response is to say the CLOCKSS and any other system that cannot be described using *only* ISO 14721 concepts and terminology is not an archival system. Whatever it is doing is not archiving. Over time, as technology and the requirements of the marketplace evolve, the terminology of ISO 14721 will describe fewer and fewer systems, so the field of archiving will shrink to encompass only legacy systems.

If, on the other hand, you believe that ISO 14721 is a reference model, your response is to say that it needs updating with additional concepts and terminology adequate to describe the systems that are doing archiving is the sense in which that word is generally used. Our experience has identified a number of areas in which updating is needed, and I hope to adress them in detail in subsequent posts. I'm sure others have found other such areas, and I hope they will address them in posts to this Wiki. Lets get to work to ensure that a revised ISO 14721 matches the reality of current archival systems. Once that is done, we will need to revise the standards based upon it, [ISO 16363](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510) and [ISO 16919](http://www.iso.org/iso/catalogue_detail.htm?csnumber=57950).

**COMMENT: OAIS does not claim to be a reference manual to design archives. It claims to:**

* **provides a framework for the understanding and increased awareness of archival concepts needed for Long Term digital information preservation and access;**
* **provides the concepts needed by non-archival organizations to be effective participants in the preservation process;**
* **provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future Archives;**
* **provides a framework for describing and comparing different Long Term Preservation strategies and techniques;**
* **provides a basis for comparing the data models of digital information preserved by Archives and for discussing how data models and the underlying information may change over time;**
* **provides a framework that may be expanded by other efforts to cover Long Term Preservation of information that is NOT in digital form (e.g., physical media and physical samples);**
* **expands consensus on the elements and processes for Long Term digital information preservation and access, and promotes a larger market which vendors can support;**
* **guides the identification and production of OAIS-related standards.**

**The last point is particularly relevant here. No one standard can cover everything. OAIS contained a roadmap for other standards needed to cover specific issues. Many of these had been addressed but the roadmap can be expanded.**