A Perspective from Digital Preservation on Sustainability of E-Infrastructures

Simon Lambert

UKRI-STFC, UK

simon.lambert@stfc.ac.uk

ORCID 0000-0001-9570-8121

Abstract – Much effort is expended on assuring sustainability of e-infrastructures and the services within them. Some of these services have a dimension of long-term preservation, in particular the infrastructure around persistent identifiers (PIDs). It is therefore worth asking whether established approaches to long-term digital preservation can offer insights into planning for sustainability. This paper examines some aspects of the OAIS Reference Model and their applicability to sustainability planning for e-infrastructures.

Keywords – E-infrastructure, Sustainability, OAIS Conference Topics – Exploring New Horizons

I. Introduction

Sustainability is a fundamental attribute of e-infrastructures. The European Open Science Cloud Declaration uses the word repeatedly as a *sine qua non* of the EOSC [1]. Sustainability means the capacity of the operations, services and activities of the e-infrastructure to continue dependably into the future. It means that users can rely on the e-infrastructure, and that the initial investment in creating it is not at risk of being wasted. Sustainability is not the same as financial viability, though often considered as such. There are other ways that sustainability can fail apart from problems with profitability or cash flow. Many smaller social network websites lost out to the irresistible march of Facebook, not because of financial unviability but because of loss of members to the more attractive alternative. The same is true for digital preservation: having the money to keep a repository or archive running, though necessary, is not sufficient—it might be that the archive's contents eventually become unusable, for reasons which are now very well known.

Some e-infrastructure components that need to be sustained also have long-term preservation aspects. Obviously, services that specifically offer long-term data storage such as EUDAT's B2SAFE [2] and Preservica's products [3] are of this type. The infrastructure for persistent identifiers (PIDs) is another example: a whole complex of interconnected access would be in jeopardy if the e-infrastructure failed, putting the record of science at risk.

Not all e-infrastructure services have this long-term dimension. A service for uploading and sharing datasets among collaborating researchers or a cloud computing offering should be sustained, but do not have such ramifications—their value is in their use at one particular time.

Given that some e-infrastructure services have a long-term dimension, it is natural to ask whether anything can be learnt from the principles and practices of digital preservation when thinking about and planning for sustainability of such services. The contention is that some concepts and models from the OAIS (Open Archival Information System) standard [4] can be related to sustainability and provide a fresh perspective on it.

OAIS is a conceptual framework, proposing a variety of concepts and models for describing the responsibilities and functions of an archive that aims to preserve information for the long term. This paper examines some of those concepts and models and relate them to sustainability in which there is a dimension of long-term preservation (as for PIDs). Clearly some of the concepts and models are highly specific to the functions and organization of an archive rather than an e-infrastructure, and so the correspondence cannot be taken too far, but nonetheless there are clear and valuable parallels.

iPRES 2019

II. BASIC CONCEPTS OF OAIS

The fundamental concept of OAIS is the Archive:
"An OAIS is an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community."

The Archive corresponds to the whole e-infrastructure, and the information that the Archive preserves corresponds to services within that e-infrastructure. Those services are made available to some community and the intention is that the services will continue to be available into the future. The e-infrastructure may not be an organization in quite the same was an Archive (though OAIS recognizes the possibility of distributed archives); the "organization" is whatever structures and mechanisms have been put in place to manage, govern and develop the e-infrastructure. The European Open Science Cloud (EOSC) has its governance structures; the EUDAT Collaborative Data Infrastructure has a network of service providers; PID service providers such as DataCite and Crossref have their business models based on membership; the FREYA project envisages a "PID Commons" for governance of the PID infrastructure [5].

Other fundamental concepts of OAIS are the Producer and Consumer. Producers "provide the information to be preserved" while Consumers "interact with OAIS services to find and acquire preserved information of interest". In e-infrastructure terms, Producers correspond to service providers and Consumers correspond to service users—whether end-users or third-party providers of other services. Table 1 summarizes the correspondence of these fundamental concepts.

TABLE 1

Correspondence of OAIS and E-Infrastructure Concepts

OAIS	E-infrastructures
Archive	E-infrastructure as a whole
Content Information (the original "target of preservation")	Services
Producer	Service provider
Consumer	Service user

III. SPECIALIZED OAIS CONCEPTS AND SUSTAINABILITY

Having established this basic mapping, it is now possible to examine some more specialized OAIS ideas to assess whether and how they can be applied to thinking about sustainability of e-infrastructures. It should be stressed again that the sustainability in question relates specifically to services with a long-term dimension; and that this is only a preliminary view, so that more thorough analysis might well throw up further analogs.

A key idea of OAIS is the Designated Community, which is defined as:

"An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time."

Thereby the Designated Community is contrasted with the wider group of potential Consumers, and the Archive has assumed an explicit responsibility to ensure that the information it holds should be understandable to the Designated Community over the long term. The relevance to sustainability is in the selection of a set of services that a certain community positively requires over the long term, and consequently the assumption of responsibility for maintaining the usability of those services. For example, the "Designated Community" (always with the long-term perspective) in the case of a PID infrastructure might be the body of researchers at large who require long-term access to the scientific knowledge base—meaning that resolution of PIDs appearing in articles, whether to other articles, datasets, software or whatever other entity must be maintained. This defines the requirement for long-term sustainability, even if the ability to mint new PIDs (of a certain type) ceases.

As an incidental note, thinking in terms of the Designated Community in OAIS terms might help to avoid the pitfalls of referring to "the community", a term which according to context (it is seldom clearly defined) might refer to distinct fields of research ("the social science community", "the particle physics

iPRES 2019

community") or merely to the set of individuals who take an interest in the functioning of the e-infrastructure (as in "engaging with the community").

A further basic idea of OAIS is that of Representation Information: "The information that maps a Data Object into more meaningful concepts," i.e. what must be provided to supplement the knowledge base of the Designated Community so that they can understand and use the information in the Archive. The idea is broad enough to encompass not only explanatory material, data dictionaries and the like but even emulation software. The analog in e-infrastructures is what must be provided to ensure continued usability of services as the environment changes. This might include new APIs or user interfaces, training offerings, best practice guides or "ambassadors" for particular domains.

The OAIS model defines a number of Functional Entities playing particular roles in the preservation endeavor. Some of these such as Ingest (accepting material from Producers into the Archive) are very specific to preservation of materials—though it might be that the activity of accepting a new service into an e-infrastructure throws up some parallels worth exploring. The two Functional Entities that do have clear analogs are labelled Administration and Preservation Planning: the former "provides the services and functions for the overall operation of the Archive system" while the latter "provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community …." This makes clear a valuable distinction between day-to-day operation of an e-infrastructure and the higher-level tasks of tracking trends in the user base and in technology. From a sustainability perspective, it should be possible for an e-infrastructure to point out how these two distinct functions are carried out.

IV.   Preservation Strategies and Sustainability

The three basic strategies of preservation implied by OAIS can be encapsulated as "hand over", "transform" and "add Representation Information" (the last including emulation as a special case).

Handing over refers to the transfer of the Archive's holdings to a successor Archive, which may be necessitated if the original Archive ceases to operate or otherwise becomes incapable of fulfilling its preservation responsibilities. It is tempting to relate this to individual services—ensuring continuity if a service provider fails—but it must be remembered that the hand-over refers to the whole Archive, equivalent to the e-infrastructure, not to an individual object within it. It would presumably correspond to a situation in which the e-infrastructure as a whole undergoes significant change, perhaps through a major shift in funding or top-level governance, and raises the question of how the "planning" and "administration" functions previously noted can be sustained in the new environment—in other words, what would the "successor" e-infrastructure look like?

Transformation in OAIS is a type of operation on the holdings that changes the content information—a typical example being a transformation from one data format to another. For the services of an e-infrastructure, this would correspond to reimplementation of the same functionality on a different platform. An interesting consideration is OAIS's idea of Transformational Information Property, sometimes called "significant property", which may be used to give assurance that information content has been preserved after a transformation (typically relating to accurate rendering of a document or image, though not limited to that). Trying to identify such properties for services of an e-infrastructure could be an avenue for focusing sustainability planning on what is essential in the long term, or what will give assurance of "authenticity" as things change over time.

The equivalent of adding Representation Information would be the case in which existing services are maintained but something has changed in the environment or user base that necessitates or at least makes desirable the additional "information". Tracking what might change and determining how to react are important functions that must be present in sustainability planning, whether the changes are in the technological base or community behavior. In fact there are two requirements: identifying what kind of thing might change in future, and monitoring to see if it actually does change. Both these are reflected in the OAIS description of the

iPRES 2019

Preservation Planning Functional Entity, which refers to "risk analysis reports, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base."

V. CONCLUSION

Examination of the concepts and models of OAIS reveals that some of them at least have correspondences in the sustainability of e-infrastructures, in some cases suggesting fresh ways of looking at and planning for sustainability, such as transferring the idea of the Designated Community or the three basic strategies for preservation. The FREYA project intends to make use of this perspective as one contribution to its work on sustainability of the global and European PID infrastructure.

REFERENCES

[1] "EOSC Declaration", Brussels, October 2017. https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

[2] EUDAT Service Catalogue: B2SAFE. https://www.eudat.eu/catalogue/B2SAFE

[3] https://preservica.com/digital-archive-software

[4] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information Systems", June 2012, https://public.ccsds.org/pubs/650x0m2.pdf

[5] https://www.project-freya.eu/en/about/mission

iPRES 2019