Repository Speed Dating

A methodology for narrowing the field

|  |  |  |
| --- | --- | --- |
| **Karin Bredenberg** | **Sven Schlarb** | **Carl Wilson** |
| *Kommunalförbundet Sydarkivera*  *Sweden*  *karin.bredenberg@sydarkivera.se*  *0000-0003-1627-2361* | *Austrian Institute of Technology*  *Austria*  *Sven.Schlarb@ait.ac.at*  *0000-0003-3717-0014* | *Open Preservation Foundation*  *United Kingdom*  *carl.wilson@openpreservation.org*  *0000-0003-1535-1770* |

**Abstract – Selecting a repository system is a task many collecting institutions have to carry out at least once. There are many challenges, while the variety of alternative systems available is a good thing, making sense of the marketplace can be difficult. Assessing potential candidates is time-consuming and it’s difficult to reuse the work of others as every organization has unique requirements. Here we present a simple methodology intended to help organizations to narrow the field by putting together a high-level set of requirements based upon the OAIS Reference Model, placed within the context of the OAIS Reference Model. This can help organizations evaluate solutions to create a shortlist of suppliers.**

**Keywords – OAIS, selecting, matrix, evaluation, working-together**

**Conference Topics – Community.**

# Introduction

Have you faced the same issues I've encountered at my institution when thinking about updating your digital repository and wanting to align with the OAIS Reference Model [REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS), ISO 14721:2012]? If you haven't yet, it's likely that you will have to soon given digital preservation's constant evolution. Once you reach that point, where do you start? How can you assess which of the variety of architectures and software systems match your criteria and requirements without immediately starting detailed and time-consuming discussions with vendors? Your organization's procurement policies might even mean contacting vendors isn't initially an available option.

Ideally, there would be a place where it's easy to compare available repository systems so you just can pick one, similar to choosing an air fryer after reading reviews by “Råd och Rön” or another consumer organization. In reality, these resources aren't available and they would likely be of limited use since organizational requirements are unique regarding the types of content preserved, institutional policies and local legislation.

In this paper, we present a methodology for carrying out a high-level evaluation of potential repository solutions or an assessment of existing repository systems against a set of requirements, we call it speed dating for repository systems. This uses the OAIS model as a device for classifying and aggregating requirements and repository features to produce an easy-to-use evaluation matrix.

# Background

One of the leading resources for working with digital repositories is the OAIS Reference Model. As the name suggests, this presents a reference model, which outlines the components for an ideal archival information system. The model’s scope extends beyond software products, covering the organization and staff that administer and manage the system. The reference model is ubiquitous in digital preservation disciplines, providing key terms and functional definitions. However, there are still very few good educational resources for the model, meaning practitioners must often educate themselves. In combination with the OAIS Reference Model, you can use the Audit and Certification of Trustworthy Digital Repositories (TDR), ISO 16363:2012 to ensure that the whole archival information system, including administrative and management functions, align with and follow OAIS. The certification standard contains a huge number of requirements, with TDR certification coming at a financial and resource cost. With the caveat that this methodology is not certifying the technical system, instead, it is certifying the whole system, including the organization that hosts and administers the repository. Other certification and self-evaluation models are available, but they still do not evaluate the system itself. Despite resources, we are still somewhat in the dark regarding the technical aspects of the digital repository itself, and what we should compare when procuring new systems or refreshing existing ones. How can we be sure that the digital system storing our information packages adheres to the OAIS Reference Model? We often have to simply trust that vendors and solution providers are implementing OAIS properly and that their system follows the model. The authors aren’t saying you shouldn’t trust vendors and solution providers. Indeed, trust in your supplier is essential. However, when it comes to choosing a product, I want to ensure it satisfies key requirements, important to my organization to make informed decisions without contacting vendors. This means that I need to be able to trust the information available online when at the information gathering stage. System descriptions, manuals and fact sheets should be open for all and easily accessible. We need sufficient accurate information to start an initial evaluation before carrying out a Request For Information (RFI) or full-blown procurement process.

The OAIS reference model is split into a number of functional areas. I want to be able to evaluate the available systems that align to these areas. I also need to find the criteria that are important to my organization, since our digital preservation mission will not be exactly the same as yours. Organizations will see different criteria as more or less important than another, meaning there are few shortcuts when evaluating digital preservation systems. It’s not possible to simply copy someone else's approach regardless of how much you like it. You need to put in the time and resources to first figure out what you want the system to do for you and then set up a matrix to assist you in making your evaluation. The goal is to be able to use detailed requirements as a guide to assist you in narrowing down the choice of available options. The labour involved can be organized in different ways, but you need to recognize that it is a necessary step in planning for a new repository platform.

# Using OAIS to Establish an Evaluation Framework

The OAIS model provides a conceptual framework for a digital archive, consisting of an organization of people and systems with responsibilities to preserve and provide access to information. The reference model includes concepts and terminology that can be used to describe and compare architectures and operations of digital archives as well as preservation strategies and techniques. The methodology described in this paper uses this framework to provide a foundation for categorizing requirements to produce an evaluation matrix for comparing archival systems. While we assume some familiarity with OAIS terminology this section provides definitions of key concepts used in the methodology.

## The OAIS Environment

The OAIS environment defines four interacting entities, producers of information, consumers of information, a management entity which sets policies for archival content, and the archive itself. The term "consumers of information" describes a broad population of users wishing to access content held in the archive. OAIS also defines a specific term for groups of consumers identified by archives, Designated Communities.

**Designated Community:** A group of consumers defined by an Archival organization by some criteria, e.g., occupation or location, who require access to particular information sets. The Designated Community may be composed of multiple user communities and its composition may change over time.

## The OAIS Information Model

The OAIS model defines an information model for digital items, known as data objects and any metadata needed to interpret data objects. These are components of Information Packages which consist of data objects and any metadata required to support long-term preservation and access bound into a logical package. OAIS identifies three types of Information Package described below.

**Submission Information Package (SIP):** An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.

**Archival Information Package (AIP):** An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.

**Dissemination Information Package (DIP):** An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.

**Designated Community:** A group of consumers defined by an Archival organization using some criteria, e.g., occupation or location, who require access to particular information sets. The Designated Community may be composed of multiple user communities and its composition may change over time.

## The OAIS Functional Entities

The OAIS reference model describes six distinct functional entities defined below, that can be used as a starting point when evaluating digital repository systems.

**Pre-Ingest/Ingest:** This functional entity represents the boundary between the archival system and incoming information packages. This means that solution coverage is more variable than for the other entities as different solutions effectively draw their own boundaries and might rely upon the task being handled by another system or in another part of the information package's creation.

**Archival Storage:** This functional entity can be regarded as the foundation of all other repository functionality. Secure long-term storage of digital content and metadata is the prime function of digital repository systems. Ingest, access and preservation planning functionality are all, in a sense, layered on top of archival storage. As such they can be improved and refined over time provided the underlying archival storage is well designed and reliable.

**Preservation Planning:** An OAIS function that encompasses archival activities required to ensure digital collections remain accessible and comprehensible over time. These activities include developing/creating strategic preservation policies applicable to all digital content, as well as any action plans specific to particular collections or technologies. This is a proactive function which identifies/anticipates changes that may impact the long-term preservation of and access to digital collections. These include internal and external changes to the archival organization, evolving standards and technologies, e.g., storage mediums or file formats, and changes in the needs and expectations of Designated Communities. This is a difficult function to automate and evaluate as much of it depends on specific institutional requirements.

**Access:** This functional entity marks another system boundary, in most cases the Designated Community are external consumers, usually researchers. The entity contains the services and functions that make the archival information holdings in the form of digital objects and related services visible to the researcher. This step involves an archivist reviewing the digital objects to make sure that there is nothing that the researcher isn't authorized to view, for example for data protection reasons. This means that a DIP for the archivist to review needs to be created before a redacted DIP for a researcher is created. This practice varies depending on your institution; sometimes the DIP is created when the ingest is made and sometimes it is created upon request.

**Data Management:** The functional entity of data management is a somewhat biased entity. It contains services and functions for populating, maintaining, and accessing a wide variety of information. This might imply that it is a database solution for handling a number of different statistics like access, billing and security control. It is also the function responsible for managing the repository’s descriptive and preservation metadata.

**Administration:** The administration entity is the catch-all for the organizational, procedural and technical glue that brings the system together, and integrates it within an organization’s business as usual activities. It has a broad scope covering non-technical processes and activities, operational management and institutional policymaking. Many of the administration functions are outside of the scope of many, if not all, repository systems. When it comes to infrastructure or process management, existing, dedicated solutions are in place. Leveraging existing domain software to address gaps in the functional coverage of the chosen solution will be the pragmatic, (possibly only) choice.

# Methodology

Organizations embarking on a procurement process can be faced with a choice between a large number of possible approaches and technologies. Attempting a thorough assessment of all available options is often time consuming and impractical. This means assessment can prove an intimidating task and lead to analysis paralysis. Here we present an overview of a methodology designed to "narrow the field" when assessing potential repository solutions. We divide the analysis into three steps that can be summarized as:

1. Gather Requirements: Define what your organization wants from a repository system.
2. Aggregate and Prioritize Requirements: Create logical groups of requirements based on OAIS functional areas.
3. Research Solutions: Carry out high-level desktop research of potential solutions and/or existing systems.

Following this approach will help institutions to classify their requirements into logical groupings consistent with the OAIS reference model. The methodology could be used for other purposes, for example to perform an assessment or gap analysis of an existing archival system.

## Gathering Requirements

The first step is to put together a set of requirements, defining what your institution wants from a digital repository. What constitutes a good requirement gathering exercise could be the subject of its own paper. Much depends on the size of your organization and the scale of your digital collection. Generally, they can be divided into two types, functional and non-functional. The distinction isn’t always clear but functional requirements describe what the system should do while non-functional requirements describe how it should do them. It can be helpful to look at requirements documents put together by other organizations, particularly organizations of similar size with similar aims. Another quick start might be the OAIS reference model itself, which describes the functional entities that comprise an ideal digital repository in some detail. These should only be used as a starting point though.

At this early stage, it can be helpful to consider system capabilities, rather than focus on overly detailed requirements. For example, the ability to scale a system to meet performance criteria might be more meaningful than trying to stipulate a maximum throughput figure that will be hard to evaluate.

## Aggregate, Group and Prioritize Requirements

Once a set of requirements has been defined, the next step is to start to gather them into logical groups, defined by the OAIS areas described previously. The initial evaluation is simply a case of researching publicly available material for compliance with requirements. Again, this is more easily performed if the requirements are more general/coarse-grained. A concrete example might be requirements around integrity checking content held in archival storage. Your organization might have detailed requirements as to digest algorithms and the number of storage nodes supported, e.g., SHA-256 checking across four archival nodes. While it’s important to note these requirements, this level of detailed evaluation will come later. For now, aggregating these together as “Audits and integrity checking” under the Archival Storage category is enough.

This simplification process might take some time. We have used six functional areas derived from the OAIS model. The goal should be to have five or six aggregated requirements per functional area. This isn’t a prescriptive rule but ten or more requirements is probably a mistake as the detailed information to evaluate them is unlikely to be available for most of the systems. To reiterate, you can reduce the number of requirements to consider by aggregating similar requirements together as a coarse-grained, more general requirement. Considering the priority of requirements is another method, low priority requirements may be left out altogether at this stage. They will be reintroduced later when carrying out a detailed analysis of the solutions shortlisted by this process.

Finally, these grouped requirements should be arranged as the top row of a matrix, in a spreadsheet. Here’s an example:

En bild som visar text

Automatiskt genererad beskrivning  
Figure 1 Matrix header row example.

## Product Research

The next step is to evaluate each of the potential solutions against the criteria listed. The aim is to identify and eliminate solutions that have obvious gaps, not to perform a detailed evaluation. The product research is carried out simply using information available on the internet such as:

* Product and vendor websites.
* Online manuals and help guides.
* Community forums if available and accessible.

For each of the requirements, only a yes/no answer is required, if the decision is difficult err on the side of generosity and give a yes. A more in-depth evaluation is the place for making trickier decisions.

As the solutions are evaluated against the criteria you simply fill in the appropriate box in the matrix, see example below. The name of the solution and the answers are randomly inserted here to illustrate the methodology.

En bild som visar bord

Automatiskt genererad beskrivning  
Figure 2 Matrix example.

# Weaknesses

This approach is useful when initially trying to make sense of the options available and eliminating those that are clearly not fit for purpose. It also leaves a reasonably objective record of the solutions considered and the reasons for elimination. It is far from a forensic investigation of detailed system specifications, that is for the full sourcing and procurement process using the solutions that remain.

Because the assessment is performed using information published by vendors it might not be possible to make an informed judgement on all criteria. If detailed information about a particular product is hard to find it may say something about the product itself or the level of support available. The results of this process will only be as good as the work put in. Ensuring that the requirements accurately reflect your institution’s priorities and taking the time to search for detailed information will improve the results.

Another potential issue is that the requirements are “framed” by the OAIS model. This means that it’s possible to overlook important criteria that don’t align with OAIS. Examples include:

* Relationship with the vendor, ensuring that the vendor is a good cultural fit for your organization.
* Institutional or national policies and regulations. While a cloud-based solution might be attractive there may be good reasons that your organization needs to control the geographic location of data.

# Next steps

This high-level evaluation is only a first step toward sourcing a digital repository system. What comes next depends upon how you use the results of the evaluation. This is still a somewhat subjective exercise, not just a case of counting yes and no scores. “Knock out” criteria can help, these are mandatory features which must be supported by candidate solutions. Beware overusing this blunt instrument, you may arrive at a situation where no available system satisfies all of your mandatory features.

One approach is to consider the next phase and how many options you can realistically evaluate. If the next phase is a full procurement, with vendor interviews and product demonstrations, then it’s probably unrealistic to consider more than 5 solutions due to the evaluation effort involved. If ten or eleven candidate solutions remain, you might decide to perform a second sift using the fine-grained requirements that were aggregated earlier. This assumes that the data needed to make more nuanced distinctions are publicly available. It may be necessary to contact vendors for more detail. This could still be more informal than an official procurement, using a questionnaire or a set of scenarios for the suppliers. In the end, it’s your decision, and approaches that work for other organizations might not be as well suited to yours.