Becoming a Trustworthy Digital Repository: Some Thoughts

Mark Phillips
UNT Libraries
July 18, 2016
This presentation discuss the process that the UNT Libraries used to complete a TRAC self-audit published in Fall 2015
Outline

Background
Process
Lessons Learned
Next Steps
Background
First a little terminology
Open Archival Information System (OAIS)
REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)

Published 2002
Updated 2012
OAIS → ISO 14721:2003

Space data and information transfer systems -- Open archival information system -- Reference model

Space data and information transfer systems --
Open archival information system (OAIS) --
Reference model
OAIS is a high level reference model
An Open Archival Information System (or OAIS) is an archive, consisting of an organization of people and systems, that has the responsibility to preserve information and make it available for a Designated Community.
The information maintained is deemed to need “long term preservation” where “long term” is long enough to be concerned with the impacts of changing technologies
Key concepts from OAIS
(according to Mark)
Designated Community
“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.”
Information Package
Figure 2-3: Information Package Concepts and Relationships
Content Information – “A set of information that is the original target of preservation”
Preservation Description Information (PDI) – “Information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context and Access Rights Information”
OAIS defines three types of Information Package
Submission Information Package (SIP)
The package that is delivered to the OAIS
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.”
The package that is stored by the 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”
The package that is delivered to users.
Trustworthy Repositories Audit & Certification (TRAC)
● Metrics of an OAIS-compliant digital repository.

● OCLC/RLG

● National Archives and Records Administration
Was revised and became.
• Trusted Digital Repository (TDR) Checklist

• Published: September 2011
Which became,
Space data and information transfer systems -- Audit and certification of trustworthy digital repositories

ISO 16363:2012
ISO 16363 standard is based upon the Trusted Digital Repository Audit Checklist (TRAC)
To complete the circle
(but not really in scope for this talk)
Once you have criteria for trusted digital repositories, you have to have some group that can audit repositories.
Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories.

Published: October 2010
Which became,
Space data and information transfer systems – Requirements for bodies providing audit and certification of candidate trustworthy digital repositories.

ISO 16919:2014
TRAC has provisions in it for a self-audit as well as an external-audit.
Process
In Fall 2014 the UNT Libraries and the George A. Smathers Libraries at the University of Florida began working on a process to each complete a TRAC self-audit.
Once each institution completed its own self-audit, the groups would critique and review the others audits, publishing the results as a form of peer-review for the TRAC process.
Fall 2014 flew by
We wrestled with questions like:
What is the “repository” in this context.
Just as a library isn't the shelves, or physical infrastructure, the repository isn't just the technology infrastructure.
It is the workflows, processes, documentation, policies, frameworks and people that utilize the technical infrastructure that make up the repository.
Who are we defining as trusted?
The Digital Projects Unit?
Digital Libraries Division?
UNT Libraries?
In the end we decided on the UNT Libraries' Digital Collections
Which is an umbrella name that covers all of the digital library activities at the UNT Libraries that use the same infrastructure/policies/workflows.
In February 2015 we really started to pickup steam once we decided on a structure for detailing our conformance to the TRAC guidelines.
What we ended up creating is a conformance document that discusses each of the 84 criteria
Each criteria of TRAC has a section and then documentation about how the UNT Libraries' Digital Collections is or isn't in compliance with the goals of TRAC.
Often times the response to a criteria will require references to other documentation that further describe the response or provide additional context.
For example criteria:
B2.1 Repository has an identifiable, written definition for each AIP or class of information preserved by the repository.
An AIP contains these key components: the primary data object to be preserved, its supporting Representation Information (format and meaning of the format elements), and the various categories of Preservation Description Information (PDI) that also need to be associated with the primary data object: Fixity, Provenance, Context, and Reference. There should be a definition of how these categories of information are bound together and/or related in such a way that they can always be found and managed within the archive.

It is merely necessary that definitions exist for each AIP, or class of AIP if there are many instances of the same type. Repositories that store a wide variety of object types may need a specific definition for each AIP they hold, but it is expected that most repositories will establish class descriptions that apply to many AIPs. It must be possible to determine which definition applies to which AIP.

While this requirement is primarily concerned with issues of identifying and binding key components of the AIP, B2.2 places more stringent conditions on the content of the key components to ensure that they are fit for the intended purpose. Separating the two criteria is important, particularly if a repository does not satisfy one of them. It is important to know whether some or all AIPs are not defined, or that the definitions exist but are not adequate.
**Evidence:** Documentation identifying each class of AIP and describing how each is implemented within the repository. Implementations may, for example, involve some combination of files, databases, and/or documents.
Statement from TRAC Conformance Document:
B2.1 Repository has an identifiable, written definition for each AIP or class of information preserved by the repository.

The Archival Information Package (AIP) used by the UNT Libraries is formally described in the UNT Libraries’ OAIS Information Package Specification (Appendix M).

The UNTL-AIP serializes the structure of the digital object with a METS record conforming to the UNTL METS Archival Information Package Profile (http://www.loc.gov/standards/mets/profiles/00000045.xml) and further encapsulates the files associated with the digital object within a bag as specified by the BagIt 0.97 Specification.
References two external documents:

UNT Libraries’ OAIS Information Package Specification(Appendix M)

UNTL METS Archival Information Package Profile
UNT Libraries OAIS Information Package Specification

October 2015

Date: October 2015
Version: 1.0
Contributors:
- Mark Phillips, Assistant Dean for Digital Libraries
- Hannah Tarver, Department Head, Digital Projects Unit
- Ana Krahmer, Supervisor, Digital Newspaper Unit
- Daniel Alemneh, Supervisor, Digital Curation Unit
- Laura Waugh, Repository Librarian for Scholarly Works

This work is licensed under a Creative Commons Attribution 4.0 International License.
UNT-AIP METS Profile registered with the Library of Congress

May 2015
Total output of the process:
1 Conformance Document
23 Appendix Documents
291 Pages of fun
TRAC is divided into three sections
Section A: Organizational Infrastructure
Section B: Digital Object Management
Section C: Technologies, Technical Infrastructure and Security
Section A covers key concepts like:

- What is your mission?
- What authorization do you have to exist?
- What permissions do you have from content creators?
- What is your staffing level in relation to your services?
- What do you do to train staff and improve skills?
- How do you budget and plan for gaps in funding?
- How do you maintain and adjust policies?
Section B covers key concepts like:

- How do you package your content?
- How do you track packages from SIP-AIP-DIP?
- How do you verify integrity of packages?
- How do you track events related to the package?
- How do you create identifiers for packages?
- How do you acquire preservation metadata for packages?
Section C covers key concepts like:

- How do you plan for technology change?
- Do you have adequate hardware, software, and support for your repositories services?
- How do you maintain multiple copies of your digital objects?
- How do you detect bit corruption or loss?
- What happens when you identify corruption or loss?
- What is your disaster recovery plan?
Locally this process looked at the following areas related to the UNT Libraries' Digital Collections.
Organization

- Mission
- Succession Planning
- Policies
- Permissions
- Staffing
- Professional Development
- Budgets
- Auditing
- Rights
Digital Object Management

• What do we preserve?
• Definition, management, and workflows for Submission Information Packages (SIP)
• Definition, management and workflows for Archival Information Packages (AIP)
• Definition, management and workflows for Dissemination Information Packages (DIP)
• Unique identifier generation, procedures and commitment.
• Coda repository
• Aubrey access system (The Portal to Texas History, UNT Digital Library, the Gateway to Oklahoma History)
• Verification and validation of Information Packages
Technologies, Technical Infrastructure, and Security

- Core technology infrastructure, servers, storage, operating systems
- Hardware and software in place for backup/replication
- How we store multiple copies of Information Packages
- How we audit and detect bit corruption
- Process for reporting data corruption or loss
- Process for managing hardware change
- Risk analysis and management
Lessons Learned
Policies can be helpful
Three library policies created
Collection Development Policy for the UNT Libraries' Digital Collections

UNT Libraries' Digital Preservation Policy Framework

UNT Libraries Digital Collections Usage and Feedback Policy

Available http://www.library.unt.edu/policies
Two others referenced
UNT Libraries' Policy on Creation and Licensing of Metadata

UNT Libraries Open Source Software Policy

Available: http://www.library.unt.edu/policies
Formal documentation takes time
Lots of time
Knowing what the end product for the work is helpful, it sets a direction and provides boundaries and limits that are sometimes needed in order to move forward.
A good environment for collaboration is important. (Google Docs)
External reviewers can be very helpful for identifying implicit knowledge of a process that needs to be made explicit.
Develop a style guide for the project.

Coda
Coda Repository
Coda repository
CODA
Next Steps
Data Seal of Approval

16 guidelines
(instead of 84 in TRAC)

Self-Audit tool
nestor Seal for
Trustworthy Digital Archives

34 Criteria

Reviewed by nestor
TRAC Team:
Mark Phillips
Hannah Tarver
Ana Krahmer
Daniel Alemneh
Laura Waugh
Huge support from:

UNT Libraries:

Administrative Offices
Facilities and Systems
External Readers:

Dorothea Salo
Nick Krabbenhoeft
Kevin Hawkins
Thank you

Questions?