Cornell University Library Repository Principles and Strategies Handbook

Prepared for the Cornell University Library Repository Executive Group by the Repository Principles subgroup, March 2018:

Erin Faulder (co-chair)
Jim DelRosso (co-chair)
Jenn Colt
Dianne Dietrich
Amy Dygert
Sarah Kennedy
Zsuzsa Koltay
Jason Kovari
Wendy Kozlowski
Chris Manly
Michelle Paolillo

Introduction

In order to better coordinate a repository ecology that includes multitudinous individual systems, and synthesize staff knowledge and expertise that spans decades, the Repository Principles subgroup of CUL's Repository Executive Group (RepoExec) has created this open handbook of repository principles and strategies.

The handbook provides support for both new and existing repository managers, comprising both recommended practices and specifically identified action steps that will allow them to track their progress and identify gaps. Each section of the handbook covers a different strategic area of repository management, standing largely on its own and linking to other sections when appropriate. Although there is no primary section order, we recommend starting with Defining **Repository Scope** and **Service Planning**.

The handbook specifically addresses principles and practices pertaining to digital repositories, where a digital repository can be defined as: a system, the purpose of which is to store, present, and preserve a collection of data for which the library provides services. That is, the term refers specifically to the application as opposed to the content (collections, objects and metadata) within.

Additionally, the handbook is designed to engender a larger conversation about repository management practices, both at Cornell and beyond. As such, it is a living document that RepoExec will continue to edit and update in response to changes in the repository landscape and feedback from readers. While the handbook points to Cornell-specific service centers for providing in-house services and consulting, it is our hope that the document may be useful to a readership beyond the Cornell University Library.

We encourage readers to provide feedback on the handbook through the live version that exists in the Cornell University Library wiki: https://confluence.cornell.edu/x/18Z0F

Below you will find the grounding principles that underlie the handbook, and repository work at CUL; the sections that comprise the handbook; and the terminology used within the handbook for the roles that should be filled when providing repository services.

Grounding Principles

1. Repositories are services that require support; they are not discrete projects or information silos; they exist within a larger Cornell repositories ecosystem.
2. Repository services are dedicated to preserving and providing access to digital objects and their metadata.
3. Repository services should be developed with users' needs, ethos, and workflows in mind.
4. Repository services are dedicated to sustainable support and access of the objects for users.

Sections

The handbook is divided into the following sections. Each section was written by a pair of primary authors, selected for their interest and expertise in the topic covered, with revisions and additional material provided by the subgroup as a whole.

Role Identification

The following terminology is used throughout the handbook to describe individuals who fulfill key roles in supporting repository services. The assignment of these roles, and even how many roles an individual may fill, can vary depending on a variety of factors. Furthermore, some of these roles will be filled within a given team or department providing repository service support, while other may be functional experts within CUL.

Once the identities of the Service Sponsor and Repository Service Owner have been determined, they should work together to identify individuals who will fulfill the other roles.

- **Service Sponsor**: Often at the senior manager level, the Service Sponsor is responsible for providing fiscal sponsorship of the repository service, as well as high-level leadership and representation to the Library Executive Group (LEG). The service sponsor may also make final determination on collection model (active or passive), as this has major implications for staffing needs.
- **Repository Service Owner**: Provides leadership on the essential aspects of operating the application, including contract negotiation and policy writing. Serves as the main point

of contact for functional concerns, including identification of appropriate format types and interoperability with other repository services. May partner with others on outreach efforts, including the Content Selector. May route concerns to technical support team, if present.

- **Repository Service Manager**: Under the guidance of the Repository Service Owner, the Repository Service Manager may be responsible for identifying and responding to technical support of the repository service, with tasks including: installation; maintenance; upgrades and migrations; and monitoring or routing break/fix responses.
- **Content Creator**: Responsible for the creation of content that is included in the repository. This role is usually independent of or external to the library and/or repository service provider. The issues surrounding this role are covered in more depth in the **Curation** section.
- **Content Selector (a.k.a. Collection Curator)**: As with the Content Creator, this role may be fulfilled by a stakeholder who is external to the Library. There may also be multiple Content Selectors for a given repository service. At minimum, the Content Selector is responsible for defining the conditions of acceptance and determining which intellectual content will be included in repository, as well as serving as the main point of contact for a collection. Depending on the collection model, the Content Selector may also define primary user communities, oversee selection and collection strategies, as well as supervise the Metadata Capturer to ensure basic metadata quality control. May provide leadership on evaluating quality of analog objects for potential digitization and ingest.
- **Metadata Strategist (a.k.a Data Modeler)**: Responsible for identifying functional requirements and selecting appropriate metadata modelling to ensure interoperability and reusability of metadata outside of its home repository. May also perform metadata maintenance and clean-up using manual as well a batch processes. It is highly recommended that you consult with CUL Metadata Services.
- **Preservation Strategist**: Responsible for consulting as needed on aspects of digital preservation and curation. It is highly recommended that you consult with Digital Scholarship & Preservation Services.
- **Rights Management Strategist**: Responsible for performing permissions analysis and consulting as needed on more advanced rights management concerns. It is highly recommended that you consult with the Copyright Information Center.
- **UX/UI Strategist**: Responsible for the design, testing, and management of the user interfaces -- both public-facing and back-end -- of the repository. They are responsible for managing accessibility audits and usability testing for these systems. It is highly recommended that you consult with Digital Scholarship & Preservation Services.

Depending on the platform, collection model, and type of content intended for the repository, other roles and responsibilities may include:

- **Metadata Capturer**: Responsible for capturing and entering required metadata as defined by the Metadata Strategist, using appropriate templates and standards as provided. This role may be fulfilled by a stakeholder who is external to the Library.

- **Digitization Team**: Responsible for working with the Content Selector to evaluate quality of analog objects for potential digitization and to then actually digitize content. It is highly recommended that you consult with Digital Consulting & Production Services (DCAPS).
- **Technical Support**: With leadership from the Repository Service Manager, the technical support team may assist with installation, maintenance, upgrades and migrations, and monitoring/routing break/fix responses. It is highly recommended that you consult with CUL IT.

Defining Repository Scope

Authors: Jim DelRosso and Erin Faulder

Introduction

Managing a repository service includes tasks that focus on defining the service, collection, and audience of the repository service. Defining the scope of a repository will facilitate the success of the service, by ensuring that the content collected will meet the needs of the intended users, and that the infrastructure will support the interaction between the users and the content. The section below scoping the content contained within the repository, the communities served by the repository, and the infrastructure the supports the repository; the issues will be discussed in more depth in subsequent sections of the handbook.

Intellectual content

Understanding what content is included in your repository clarifies the boundaries of your repository service. The collecting scope could be discipline or topically specific, format specific, or audience specific. A clear collecting focus will help shape consequent decisions around infrastructure, metadata needs, copyright concerns, preservation strategies, outreach, and service planning.

- ☐ Identify content to be included in collection
- ☐ Identify content that could be excluded from collection.
- ☐ Identify any mandates or existing guidelines for content to be included in your collection.
- ☐ Write a collecting policy that describes the content your repository does and does not collect.

Content formats

Different repository systems and structures store and deliver different sorts of content with various levels of utility: whether using an existing system or starting from scratch, you need to make sure the provided functionality matches your (and your audience's) needs. Examples of different content types include documents, images, video, multimedia exhibits, etc. It's important to note whether you will need to deliver multiple kinds of formats through the same system.

RepoExec can help by working with you to determine whether any extant system(s) already house and/or deliver the kind of content you're working with. If such a system exists, you may be able to incorporate your content into it. If there is no extant internal system that meets the needs of you and your audience, RepoExec can help you find one that does.

- ☐ Identify the range of format types to be included in the repository.

- ☐ Use the CUL Repository Inventory, or consultation with the Repository Executive Group, to determine whether these content types are already delivered through an existing system.
- ☐ Contact the Repository Executive Group to identify and evaluate systems which can handle the content types you have identified, and best deliver that content to your audience.

Active and passive content collection

The two broad categories of collection for a repository are active collecting and passive collection. Active collecting involves repository staff identifying, acquiring, preparing, and uploading (ingesting) material to the repository. Passive collection has that work done by individuals or groups external to the repository staff.

- ☐ Determine whether your repository will engage in active collecting, passive collection, or a combination.

If your repository will engage with active collecting:

- ☐ Identify the staff responsible for each stage of the process: identification, acquisition, preparation, and upload (ingest).
- ☐ Determine workflow(s) for this process.

If your repository will engage with passive collection:

- ☐ Determine what community(ies) will support your passive collection. This may be the same as or differ from your access community(ies).
- ☐ Determine how repository service will engage with identified communities. Complete the **Outreach** section of the handbook.
- ☐ Identify the systems that will allow external actors to add material to the repository.
- ☐ Determine the safeguards necessary to vet content from external actors.

User communities

Defining your primary community(ies) who will be accessing the content helps guide infrastructure and development decisions. Even if the content may be available to everyone, there are still targeted audiences for the content and functions of the repository. (Everybody is not a good answer.)

One of the main purposes of establishing a digital repository is to facilitate broad access to content, so it is important to determine who you are trying to reach. Some issues to consider:

- Does access need to be restricted to a certain community? If so, some system of access control needs to be included in your repository system. Common restrictions include logins, IP restrictions, or limitation to use by the Cornell community. Sometimes only parts of your collection will require such restrictions under certain conditions. Make this determination, and then make sure that this functionality is supported by any repository

technology you choose. There are two general approaches to managing access. First, the content is considered open unless there is a reason to for it to be restricted. Second, the content is considered restricted unless certain conditions allow it to be open. Different repository infrastructures are structured to make one of these approaches more viable than the other.

- What use cases do you anticipate for this content? It's important to identify clearly how you see your audience interacting with this content; download and/or viewing are the most common, but not the only possibilities. Make a list of the different interactions your audience may have with the content, and use that to evaluate the systems you're looking at.
- Is this audience already served by an existing repository? If there's already a repository at CUL that serves this audience, it's worth examining it's functionality in light of the other issues on this list. We have a lot of repositories at Cornell, and there is no need to reinvent the wheel if you can use a system that is already in place.

- ☐ Identify designated community for content in repository.
- ☐ Identify whether designated community is supported by an existing repository within Cornell.
- ☐ Identify whether designated community is supported by an existing repository external to Cornell.
- ☐ Identify types of access controls (what does this even mean?) based on content and designated community.
- ☐ Identify anticipated use of content based on designated community(ies).
- ☐ Identify potential ways the community(ies) may interact with the content.
- ☐ Prioritize the potential interactions with content to help guide development.
- ☐ Communicate with designated community(ies) as needed to answer the above questions.

Assessment of existing repositories

Cornell already has a number of repositories; there are also many external subject repositories. Identifying gaps in collecting areas at other repositories may be an opportunity to talk with a repository that could easily fill the collections gap or provide a slightly broader range of services rather than building an entirely new repository.

If existing repositories are unable to fill the gap, articulate your collecting scope within this gap as part of your repository's mission.

If there is overlap between collections and communities, it may save you on development time and money to participate in an existing repository that may require small modifications to support your collections and community(ies).

- ☐ Survey repositories for those with similar collections.
- ☐ Survey repositories for those serving similar communities or audiences.

- ☐ Determine if an existing repository serves all or most collections and audiences. If so reach out to repository service manager and have a conversation about expanding scope of existing service.
- ☐ Consider cost/benefit of starting new service vs. expanding current service.

Interoperability

It is also important to consider how your repository will interact with other services, both those maintained by Cornell and those beyond. Many of these issues will be covered in other sections of the handbook, but it is important to begin thinking about them now.

For example: how will your audience discover the material in your repository? Will they use CUL's discovery layer, Google, both, something else entirely? How will the content you've been entrusted with be preserved? Will the repository have a standard means of displaying content to your audience, or will you need specialized systems or exhibits for some collections? All of these questions will necessitate certain design and implementation decisions, and will often involve working with established groups within CUL.

- ☐ Complete the **Preservation** section of this handbook.
- ☐ Complete the **Infrastructure and Interoperability** section of this handbook.
- ☐ Complete the **Access: Discovery and Delivery** section of this handbook.

Service Planning

Authors: Sarah Kennedy and Jim DelRosso

Introduction

Repository service planning involves taking the necessary steps to ensure the long-term sustainability and stability of the repository. Service planning may address concerns that are both functional and personnel-related, including such tasks as: staffing and resource assessment, role identification, policy writing, contract renewal and negotiation, and succession planning.

Needs Assessment

Whether you are starting a repository service from scratch or considering making upgrades to an existing repository service, it is vital to assess the needs of your community. Without knowing what needs the repository service will fill, it will be nearly impossible to design a service model that will prove effective.

If you are already providing repository services, having a clear sense of why you are providing them and how well they are meeting your current goals and expectations may provide insight and direction for resource planning and staffing, future development, and potentially also a sense of the skills and competencies that you need to hire for in the future.

In either case, it is important to look at the existing repository landscape at Cornell, and determine whether an independent infrastructure will best fit the service needs of your community, or whether an existing system will be better suited.

The steps of this needs assessment can be found below; we also recommend working through the Repository Scope section of this handbook as part of that analysis. Semi-structured stakeholder interviews may be a good way to collect some of this information.

- ☐ Complete the **Repository Scope** section of this handbook.
    - ☐ Evaluate whether your audience and content can be served by an existing repository at Cornell.
    - ☐ Work with stakeholders to identify functionalities that you would wish for in your ideal repository service.
- ☐ Consult with your Preservation Strategist to ensure long-term preservation management. (For more detail on this bullet point, see the section on **Preservation**.)
- ☐ If you're providing repository services already, document what works well with your current repository service and what could work better.
- ☐ Perform a needs assessment to determine whether or not your current collection model is serving the needs of your constituents.
- ☐ Identify stakeholders that you would like to reach but have not been successful in doing so.

- ☐ If you currently have insufficient staffing and resources to complete your work to the highest possible standard, describe additional staffing and resource needs.
- ☐ Document both your process and your results.

Staffing and Resource Planning

Whether you are setting up an entirely new repository, migrating extant content to a new platform, or undergoing efforts to scale up ingest of content within your current platform, it is important to have at least a rudimentary tool or checklist to generate an estimate of your staffing and resource needs.

Given the variances in repository platforms (open-source vs. proprietary), content types, and collection and staffing models, however, it is difficult to design a one-size-fits-all checklist. Instead, here are a few action items that you might consider as you generate a rough estimate of effort (in a 12-month time frame) and also a few actions to consider if you plan to scale up ingest. It is highly recommended that you also consult the section on **Defining Repository Scope**, which goes into much greater detail regarding the action items below.

- ☐ For purposes of comparison/benchmarking, identify similar efforts that may have taken place within other (Library) units. Consult the Repository Inventory produced by the Repository Inventory Working Group.
- ☐ Identify ingest type (e.g. self-deposit, mediated deposit, batches, or individual file uploads). See **Defining Repository Scope** section, specifically the "Active and passive content collection" portion.
- ☐ Document the number, type, and size of files you or external partners uploaded to your repository in the past 12 months. Extrapolate growth projection of annual additional records and file size.
- ☐ Calculate roughly how many hours per week were dedicated to this effort and by whom, including staff assistance from other units (e.g. eCommons Administrators team). Extrapolate growth projection of annual FTE.
- ☐ Consider whether you need to digitize the collection material. See **Curation** section for more details.
- ☐ Review new or desirable functionalities gathered from stakeholders during your Needs Assessment, and consider whether these new functionalities will require in-house development. If so, consult with CUL IT.

Role Identification

A critical step in planning for a successful and sustainable repository service is developing a shared understanding of the core roles and responsibilities involved in managing your repository service. A clear understanding of who is accountable for which actions (and when) eliminates ambiguity and promotes a sense of shared as well as individual responsibility. It should be noted that roles and responsibilities may differ slightly depending on your platform, collection model, and type of content. Additionally, the degree to which these roles are consolidated or singular may depend on the scale and collection model of the repository. It is

also worth noting that many of these roles can and should be filled by experts within CUL, rather than in your department and team.

The roles used in this handbook are described in the **Introduction**. As noted there, once the identities of the *Service Sponsor* and *Repository Service Owner* have been determined, they should work together to identify individuals who will fulfill the other roles.

Succession Planning

Succession planning, or the preparation for the departure of key personnel with institutional memory, is a multifaceted process critical to the long-term success and sustainability of a repository service. At a minimum, succession planning involves the capture of key institutional knowledge from the departing staff person as well as the identification and grooming of a suitable successor. A more programmatic approach to succession planning, however, may also include developing a clear understanding of how the repository service fits into the broader repository landscape and strategic goals of the institution.

As you prepare for a major change in staffing, then, it may be an opportune time to pause and reflect on the role of your repository service in the context of broader institutional changes and priorities. It might be useful for repository managers to consider the following guiding questions:

- ☐ Where have we been?
- ☐ Where are we now?
- ☐ Where are we going?

Where have we been? Or, what institutional knowledge must be captured as longtime staff depart their roles?

Knowledge management, or the capturing and documentation of institutional knowledge, is a core aspect of succession planning. Here are a few action items to consider as you begin a retrospective look at your repository service and interact with departing staff persons.

- ☐ Document the initial impetus for starting the repository service, identifying the initial need that it fulfilled and the intended audience.
- ☐ Identify the initial collection focus and whether that focus has changed over time.
- ☐ Locate extant workflows, standards, policies, contracts, or other documentation maintained by previous Repository Mangers through time. These may or may not be written down.
- ☐ Identify the historical collection model (i.e. active or passive), and whether this has changed over time.
- ☐ Identify access credentials that must be passed to future repository managers.
- ☐ Document intellectual stewardship of individual collections and plan to (re)assign stewardship as needed through retirements and institutional change. (For more detail on this action item, see the **Curation** and **Rights Management** sections.)

Where are we now?

- ☐ Consult the sub-section on Needs Assessment of the Service Planning section.
- ☐ Consult the sub-section on Staffing and Resource Planning of the Service Planning section.
- ☐ Consult the **Assessment of Impact** section.

Where are we going? Or, will the repository continue and, if so, how?

- ☐ Consult with the Service Sponsor to determine whether or not the repository will continue. If not, skip to "Exit Planning". If so, complete action items below.
- ☐ Revisit the job ad or position description for the new Repository Manager. Update the required and preferred skills as needed, accounting for changes in technologies, core competencies, and/or workload.
- ☐ Identify aspirational skills or competencies to consider for the successful candidate's growth and development. In other words, make allowances for robust professional development opportunities.
- ☐ Identify areas for potential growth and expansion (e.g. new audience(s), expanded content types, etc.) See the sub-section on Staffing and Resource Planning of the Service Planning section.
- ☐ If you do not already have one, consider writing an outreach and/or communication plan in order to maintain and build partnerships with key stakeholders. See the **Outreach** section.

Contract Negotiation and Renewal

If you're using licensed software for your repository, contract negotiation and renewal will be a key part of ensuring the provision of services. Initial negotiations should involve the Service Sponsor and Repository Service Owner, in consultation with University Counsel and the Copyright Office. Other roles may be consulted as needed, especially the Metadata, Preservation, Rights Management, and UI/UX Strategists.

A helpful way to frame this negotiation is, *What roles from the list above will be filled by the software licensor?* Most commonly, the licensor will be responsible for the tasks and duties associated with the role of Repository Service Manager, but if this is not the case, those duties will need to be assigned internally and the internal Repository Service Manager will need to able to perform them within the licensed software. In any case, such an arrangement should be reflected in the contract to the satisfaction of all parties. The same is true if there are any other roles that the licensor will be called on to fill.

It is also vital to take into account the policies noted above: can they be fulfilled under the contract as written? For this reason alone, it's a good idea to have those policies in writing before negotiations!

The Service Sponsor, University Counsel, and Copyright Office will likely have their own concerns about the contract, as well. Only when all of these concerns have been answered should the contract be signed.

All of these issues should be revisited when the time to renew the contract comes around, to see how well the service has fulfilled its requirements to date. Renewals afford us the opportunity to evaluate the system as executed, and thus make necessary improvements. It's a good idea to expand the people involved in the conversation to include other roles, who may have important insights into what improvements can be made, or whether the contract should be renewed at all.

Note that while this section largely deals with licensed repository software, the principles involved can be applied to homegrown or internally managed open source software as well. It's common in such cases that the roles above will be filled by colleagues from different departments, and internal MOUs can help insure sustainable service, even through reorganizations and staff changes.

If working with an external vendor:

- ☐ Work with stakeholders above to ensure that the language of the contract meets all repository needs.
- ☐ Get sign-off from Service Sponsor, Copyright Office, Budget office, and University Counsel.
- ☐ Distribute copies of signed contract to all relevant stakeholders.
- ☐ Review contract yearly to see if changes are necessary.

If working with an internal unit:

- ☐ Compose a written agreement detailing each unit's responsibilities and obligations.
- ☐ Work with stakeholders above to ensure that the language of the agreement meets all repository needs.
- ☐ Get sign-off from Service Sponsor, Copyright Office, Budget office, and University Counsel, as well as the leadership of the unit you're working with.
- ☐ Distribute copies of signed agreement to all relevant stakeholders.
- ☐ Review agreement yearly to see if changes are necessary.

Exit Planning

Exit planning is the preparation to discontinue a repository service and, like succession planning, is not something which should be postponed until circumstances demand its implementation.

It includes the logistical and infrastructural details of shutting down a repository service and migrating historical content from one repository platform to another. While the details of the new repository platform are almost certainly unknowns in the early part of the process, steps

can and should be taken in preparation of the inevitable transition. If your repository is built on a licensed platform, succession planning may also include negotiating contract termination: see Contract Negotiation and Renewal above, and consider the elements of your contract in the context of how the relationship it describes will end.

Such preparation is necessary because, while an exit plan may be deployed after a pre-determined trigger (e.g. depletion of absolute funds or foreseen & agreed upon contract terminus), in most situations it will be difficult or impossible to predict when an exit plan will need to be deployed. Sudden mergers and acquisitions among licensed platforms, lack of responsiveness or support from open-source development community, lack of resources or expertise to support in-house development of local open-source installation, or a movement from a single-institution to a shared or multi-institution solution may all put you in a position such that enacting your exit plan is necessary. Better to have it prepared before that moment.

If and when you find yourself in the position of discontinuing a repository service and migrating historical content to a new platform, **we recommend that you read through the entirety of this handbook to cover all bases.** Much of what has been covered will need to be revisited, if not established for the first time.

At minimum, complete the action items below:

If Exit Planning from Licensed Platform

Exit planning from a licensed platform is more complicated than exit planning from a local, open-source platform for 2 main reasons: (1) data restitution and (2) contract termination.

- ☐ Familiarize yourself with the conditions under which a contract can be terminated, with especial notice of the timeframe surrounding notification of parties and the end of services.
- ☐ Consult with University Counsel and Service Sponsors to confirm adherence to your contract.
- ☐ Ensure complete restitution of data, including both the content itself as well as descriptive metadata.
- ☐ Ensure that restituted data is useful (i.e. usable and interoperable).
- ☐ Once all data has been successfully migrated (see below), ensure that the original provider destroys every copy of the data on their data centers.

If Exit Planning from a Non-Licensed Platform

While leaving a non-licensed platform should remove the concern around contract termination, other issues remain.

- ☐ Ensure that all data in the current system is useful -- usable and interoperable -- in preparation for migration (see below).

- ☐ Work with current service providers within your unit to establish procedures for getting that data, and deleting it post-migrations.

For All Exit Planning

- ☐ Document institutional knowledge. See the sub-section on Succession Planning, specifically "Where have we been?"
- ☐ Identify the new infrastructure by which repository services -- including access to the extant content -- will be provided. **This will involve working through this entire section of the handbook, and likely others, completely from scratch.**
- ☐ Work with Repository Service Owner, Repository Service Manager, Metadata Strategist, Preservation Strategist, and other roles as needed to create a migration plan for extant content.
- ☐ Consult with sponsors, content creators, and other stakeholders regarding the change of infrastructure, any differences to the level of service that they may experience, as well as any interruption of service during migration.
- ☐ Establish training time for any team members working with the new software.
- ☐ Document the entire process; you will need to do this again.

Access: Discovery and Delivery

Authors: Erin Faulder and Dianne Dietrich

Introduction

Repository access is about providing end users with material that serves their needs. Users in access systems can be humans navigating a GUI environment, humans querying the repository using machines through backend tools (API queries), or machines scraping the repository (API calls, OAI-PMH). Repositories should consider to what extent they allow access of content to multiple types of users.

Access is comprised of two functions: Discovery and Delivery.

- **Discovery**: Methods allowing user to find the information object that answers the question they were asking. Discovery services can use a combination of metadata and extracted content (e.g. OCR'd documents, audio transcripts) to answer the user's query.
- **Delivery**: Serving up metadata and digital object(s) as requested by the user. Delivery methods can be browser rendering, downloadable files, metadata returns.

Discovery

Discovery refers to the discovery of content from both within and outside the repository. Is your content discoverable from the open web? How can your content be searched from within by the repository's search functionality?

Discovery from the open web

Discovery in this sense is closely tied to the best practices of metadata and interoperability. The exposure of particular metadata fields and their representation greatly affects the ability of users (and machines) on the open web to find your content.

- ☐ Consult with metadata services about metadata standards designed specifically for open web discovery, like Schema.org.
- ☐ Work with developers to ensure that the markup of your pages includes tags and semantic markup that will improve the search engine discoverability of your content.

A documented, dependable API will make it easier for machine users on the web to consume or search your content. These users may be search engines, but they could also be researchers who want perform text analysis or reuse your content in other ways. API consumers can also be bad actors, overloading your server or using your content for undesirable purposes.

- ☐ Work with your designers and developers to create API access that will best serve your users and their needs.
- ☐ Document API access.

Discovery within your repository

Discovery within your repository should meet your users' needs and expectations as well as possible. Two kinds of data are important for discovery - assigned metadata and extracted content. Full-text search tends to be an important use case for repositories but extracted content can also be used to search images, audio visual material, data files, and other types of content. It is important to carefully understand the types of discovery that users need and expect for your content. Implementing full-text search for large and complex objects is non-trivial.

In addition to configuration related to extracted content, a number of other discovery options and configurations should be considered:

- Fields that can be sortable
- Facets based on metadata
- Metadata that can be clicked on to prompt a new search
- Discovery of relationships between objects (e.g. content in same collection) and within objects (e.g. pages within a book)
- Browse functionality
- Shareability of specific content based on URL construction
- Ad hoc grouping of content into collections
- Controlling levels of discoverability for access-controlled objects, including withdrawn objects (see mediating access below)
- Language support, including language-based configurations for stemming, spell checking and synonym dictionaries

- ☐ Identify discovery functionalities needs for your users.
- ☐ Identify discovery methods for repository content.
- ☐ Evaluate the configurability of your solution for accommodating new discovery requirements as they arise.
- ☐ Test discovery mechanisms for accessibility compliance.

Delivery

Delivery is the method by which content discovered by user is served up to them. In a repository, delivery of content is contingent on file format(s), access and use restrictions, browser type, digital object size. Web delivery presents challenges due to the variability of environments the user may use to interact with the repository and its content. In order to grapple with these issues, there are two approaches to delivering content discovered within a repository.

- **Browser rendering of content**: users will be able to see and interact with the digital content without needing to download a file to view on their local computer.

- **Downloading content**: users download content to use material in an environment outside of the repository. Burden of rendering the content is on user and their computer environment.

Often there may be multiple delivery options for objects in a repository, and there are issues that affect delivery methods:

Format being delivered

Certain formats are more easily rendered in a browser than others. Certain formats researchers may want to work with outside the repository environment.

- ☐ Identify formats being delivered.
- ☐ Determine preferred delivery option for formats in repository.

Accessibility

Repositories should deliver content in an accessible way. This may mean delivering transcripts alongside video that are time coded together or ensuring large objects that are delivered through download are clearly identified with recommended connection capabilities.

- ☐ Identify accessibility requirements for formats and their preferred delivery method.
- ☐ Identify what kinds of technical platforms are and are not supported (computer browser, mobile applications, etc.)

Aggregate vs. single item delivery

Decisions about delivering a single item or aggregate objects will affect your data model. A single item may be easier to render in a browser within a repository than an aggregate object (e.g. a .zip file, .xsl file).

- ☐ Determine whether content can or should be delivered as single item or aggregate.

Delivery restrictions

Digital content may have restrictions that impact delivery options. These may be:

- Time specific (e.g. embargo)
- Location specific (e.g. only certain countries)
- Person or role specific (e.g. creator only; Cornell community only)
- Content use or reuse allowances (e.g. embed streaming audio or video on another website; downloading entire collection; downloading only low-resolution derivatives)

- ☐ Identify content restrictions that may impact delivery options.
- ☐ Document the types of restrictions on content that affect delivery.

Mediating access

Often content in repository may have limitations on access. Limitations may be on how content is discovered, when users can get content delivered, or even what types of users may get any access to the content. Mediation is managed through policy and authentication.

Policy

Mediating access through policy means setting rules governing acceptance of content into the repository (e.g. only open access content; no embargoes allowed; all content is available to certain designated community but not the open public). Policies help limit the complexity of access mediation mechanisms repository infrastructure requires.

- ☐ Establish policies defining supported access conditions for content.

Repository authentication

Repository authentication can mediate access to content based on policy, user identity, metadata about content, and rules defining logic to provide appropriate machine-actionable access. Authentication could be passive (e.g. IP restriction) or active (e.g. user login).

- ☐ Identify user type(s).
- ☐ Identify discovery restriction(s) for content in repository (who can discover what how?)
- ☐ Identify delivery restriction(s) for content in repository (who can get what content how?)
- ☐ Determine what policies govern access mediation.
- ☐ Determine what level of authentication is needed for content.
- ☐ Document rules needed to provide appropriate types of discovery and delivery to user(s).

Personnel cost to mediating access

Mediated access may also add an administrative burden to properly identify content requiring mediation and delivering content through a potentially more manual process. Self-serve access may require a more mature repository infrastructure.

- ☐ Identify conditions requiring mediated access.
- ☐ Evaluate potential personnel cost necessary to provide timely mediated access given demand of material.

User experience

The user experience of your repository is, by its nature, informed by many of the other principles described in this document. In particular, the user stories and features described during the initial definition of your repository and the metadata that powers your user experience are essential to providing discovery and delivery functionality to your users.

Successful discovery and delivery requires that you undertake the best practices of metadata outlined in elsewhere in this document and combine them with best practices from user experience research and design.

User experience practices to undertake include:

- ☐ Define your user community and then collaborate with user experience, research, and usability groups within CUL to develop personas, use cases and other UX deliverables to communicate your understanding of users and their needs to repository designers, developers, and maintainers.
- ☐ Conduct ccessibility reviews of repository solutions and necessary remediation throughout the repository's life in service.
- ☐ Conduct usability testing and accessibility reviews to verify that repository design and development decisions meet your users' needs throughout the repository's life.
- ☐ Review usage statistics regularly to answer questions about the ways your audience is using the repository.
- ☐ Conform to best practices of interface design and maintenance. An appropriate, modern look and feel with clear institutional associations promotes user confidence and greater use of the repository. Repository solutions should always include the capacity to display appropriate Cornell University Library branding.

Accessibility

In order to provide equitable access, accessibility should be a primary consideration in repository access. Accessibility considerations include:

- Vision impairment
- Auditory impairment
- Dexterity impairment
- Language of user
- Cognitive abilities of user
- Mobile technology
- Internet speed

Cornell University Library works toward meeting a minimum of WCAG 2.0 AA Standards. Any new web sites or repositories should meet this standard. The CUL User Experience team can advise you on accessibility best practices and connect you to other Cornell University resources. Facilities for testing access for disabled users are available on campus through student disability services. In addition, they offer regular trainings for web developers and others providing online resources to disabled students. Cornell IT now also offers a University-wide Site Improve license which provides automated accessibility testing.

- ☐ Identify accessibility minimums for repository based on user communities.
- ☐ Talk to CUL User Experience Team.

- ☐ Be sure that your development team makes accessibility checks part of the development process and not an after thought. Many tools exist to help them with this task. Tools/checklists to consider: A11Y CLI Audit Tool, WCAG 2.0 QuickRef, the A11Y Project's checklist, WAVE Browser Extensions.
- ☐ Use university supported captioning for A/V collections: https://it.cornell.edu/vod/captions-recommended-method
- ☐ Implement accessibility testing criteria.

Usability studies

Usability studies help evaluate whether users are able to perform the tasks required to interact with content in your repository. These studies provide a process to examine ease of use, intuitiveness, learnability and overall satisfaction that users experience when interacting with the repository. Repository systems are more complex than many other library websites because users may be both creating and consuming content, as well as sharing content, updating content and mediating the access of others to the content. All of these types of interactions should undergo usability testing. You may also have staff spending many hours working on the repository, testing for efficiency in their UX may be important to you as well.

CUL and Cornell have support in place for conducting user and usability studies. A good first step is to contact CUL's Usability Working Group and CUL's Research and Assessment department. These groups can help you define testing goals and implementations.

Initial development

The studies you conduct will be determined in part by your current development phase. For instance, early usability studies should involve low fidelity prototypes that demonstrate the in-progress nature of the repository's development, granting users the freedom to offer wide ranging feedback without fearing they will derail or insult your project. Later studies could include ethnographic observation of users conducting real research using your repository and its resources. Studies are not limited to one type of test or the collection of only one type of data.

Ongoing assessment

In addition to the data collected in formal studies, you will be collecting data from your users on an ongoing basis through use statistics, user feedback, and errors encountered by users. At the outset, do your best to provision your repository to be able to respond to this data. This includes setting up reliable feedback mechanisms users can employ and determining how you will respond to feedback. Having both a developer and designer assigned to support your repository will help you determine how and when to implement changes related to the feedback you receive. Tracking and responding to feedback data will fuel the ongoing improvement of your repository.

- ☐ Conduct user studies and usability tests throughout the life of your repository using appropriate methodology.
- ☐ Track and analyze passively collected data from your users.
- ☐ Implement a plan for the processing of feedback into a sustainable design and development cycle.

Assessment of Impact

Authors: Jim DelRosso and Zsuzsa Koltay

Introduction

This section discusses the assessment of repositories' impact on users. While this is usually the connotation of the word "assessment" when it's used in a library context, it should be noted that there are many other aspects of repositories that should be assessed and evaluated, including:

- **Curation**
- **Metadata**
- **Software infrastructure**
- **Documentation**
- **Preservation**

The sections noted above address the methods of assessment appropriate to their focus.

Similarly, while the connotation of the term "user" is often "end user," this section does not speak exclusively to that cohort. The practices detailed can also be applied to other user groups; for example, unit librarians responsible for recruiting content into a repository (see: http://blogs.cornell.edu/dsps/2017/10/30/scwg-ecommons-and-unit-libraries/).

Repository managers can assess the impact of repositories in a number of ways depending on the primary goals of the project, current concerns, and research questions. Assessment results can be used as evidence of success, for outreach, publicity efforts, reporting, improvement of service, or for internal evaluations of workflows. Visualizations of that data such as maps, graphs of trends, or even interactive tools can enhance those efforts by facilitating stronger narratives.

Developing an assessment plan should not wait until stakeholders ask you for this information. Fortunately, Library Assessment & Communication can help you think through your needs and appropriate approaches, and can be reached at researchandassessment@cornell.edu.

Usage statistics

Depending on the purpose of your repository, there may be different types of use -- and therefore, different kinds of usage statistics -- that you will want to track. While this section largely focuses on repositories with a strong dissemination function, a repository whose purpose is primarily preservation (for example) will likely measure its impact in different ways. Be sure to identify the use that is most important to assessing your impact, and pursue statistics appropriately.

If your repository does serve a dissemination function, then its software should be able to provide information on how many items have been added to the repository, as well as how many times items in the repository have been downloaded (or otherwise accessed), and the Repository Service Owner and Repository Service Manager should familiarize themselves with how to access and report on this data. The decision of whether or not to include unintentional or automated uses (e.g. double-clicking or access by robots) may vary. Whatever the decision, it is best practice to publicly document how your repository addresses (or is unable to address) such common data issues.

Beyond the reporting features of the native systems, third party analytics applications can also be used to collect usage statistics for public-facing websites, including repositories. CULIT maintains a central instance of Piwik which can be used for this purpose; to get started with Piwik for your repository, contact library-systems@cornell.edu. For assistance in reading your reports, contact libraryux@cornell.edu.

Assessment and Communication collects certain usage statistics annually, and reports them out to national agencies; items held and download counts for institutional repositories are currently reported to ACRL, for example. Cornell University Library also uses these figures for internal trend tracking: see the collections and collections use table of the CUL Statistical Trends report.

Both native and third-party systems can track other sorts of use information which may prove helpful to your assessment efforts: e.g. where users get referred from, users' domains and geographic locations, Cornell vs. non-Cornell use, and unique user data that allows for calculation of the percentage of a user population uploading material, etc. Some systems also track downloads to the article or collection level. These data might be useful to answering specific questions, so it's important to talk with your stakeholders to identify their needs. While these data are not generally collected by Library Assessment & Communication, they may contact you if the needs are identified on their ends.

Usage data collection also raises questions of user privacy. Legally and ethically, we cannot share personally identifiable information about how any individual patron uses our collections, therefore we must be vigilant about anonymizing the data and about our processes and assessment methods that could lead to the identification of a patron's use of our materials, deliberately or inadvertently. It is possible for the combination of data from multiple sources to identify a specific patron's activities, even if individual data sources cannot do this alone. Be sure to work with Library Assessment & Communication, CULIT, and the Director of Copyright as appropriate to make sure that your methods adhere to our policies and the law.

- ☐ Type your task here, using "@" to assign to a user and "//" to select a due date
- ☐ Identify the kind of usage statistics that best reflect the impact of your repository.

- ☐ Identify and familiarize yourself with the native use assessment systems of the repository software you are using, or considering using, to ensure that you are able to access the usage statistics you've identified as important.
    - ☐ If appropriate, determine whether and/or how to track unintentional or automatic usage, and document your decision.
- ☐ Consult with CULIT to identify and familiarize yourself with any third-party systems in use at CUL that might also be implemented to assess repository use.
- ☐ Contact Library Assessment & Communications regarding their recurrent usage data needs.
- ☐ Identify and consult with other stakeholders with whom you will share usage data.
- ☐ Consult with Library Assessment & Communications about other forms of data collection, as well as CULIT and the Director of Copyright as appropriate to ensure adherence to policies and laws.
- ☐ Familiarize yourself with the privacy and confidentiality policies of the Cornell University Library.

Impact stories

Impact stories relate how the broader visibility of material in your repository influenced your users, through events such as serendipitous discoveries, new partnerships, or career opportunities. These stories are the step beyond usage stats, tracing not just how often patrons access material, but how that material changes and shapes them, and how it is shaped by them in turn. This demonstrates the value of the repository in a more resonant way.

Impact stories can come to you unsolicited from satisfied patrons, so you'll need to work with the patrons to determine whether the stories can be shared publicly, and under what restrictions (e.g. anonymity). Library Assessment & Communication can help you find ways to leverage shareable stories for broader publicity, as well as solicit stories more systematically through qualitative studies.

While we cannot (and should not) track any individual patron's use of our repositories, the stories our patrons are willing to share can help define the value of our repositories.

- ☐ Track user stories that come to you unsolicited.
- ☐ Identify patrons who may have good stories to tell.
- ☐ Confirm your patrons' permission to share their stories and be publicly quoted
- ☐ Pass quotes on to Library Assessment & Communication at libcomm@cornell.edu who maintain a database of testimonials and can consult on publicity.
- ☐ Consult with Library Assessment & Communication about qualitative user studies.

User satisfaction

User satisfaction can be assessed on either a cyclical or an ad hoc basis, depending on your resources and needs. The most common means of assessing user satisfaction is via surveys, which can be disseminated on the system itself to reach users directly, or through email or

other outreach means. The latter has the benefit of potentially interacting with non-users as well, and determining why they are not accessing the repository.

Use of repositories has been included in broad-based CUL user surveys, for example frequency of repository use in the 2016 graduate student survey. Contact Library Assessment & Communication to find out what such surveys can tell you about repositories, or how to best create and deploy a survey customized to your needs.

- ☐ Determine whether cyclical or ad hoc user satisfaction surveys meet the needs of your repository service.
- ☐ If cyclical, establish your assessment schedule.
    - ☐ Contact Library Assessment & Communication to see if any existing survey results contain relevant information.
    - ☐ Contact Library Assessment & Communication to see if any of their cyclical surveys can be expanded to include questions about your repository.
    - ☐ Contact Library Assessment & Communications for a consultation on creating and deploying a customized user satisfaction survey.

Usability

Usability testing assesses how well the system interface of your repository and the organization of its content work in terms of ease and intuitiveness of use. This process is described in more detail in the **Access: Discovery and Delivery** section of this handbook.

Curation

Authors: Michelle Paollilo, Jim DelRosso, and Erin Faulder

Introduction

Content within a repository requires selection and management over time in order to support the identified communities. This work includes developing relationships with the appropriate content stewards and determining whether the content needs to be created from analog originals or if it exists in an appropriate digital format already.

Intellectual stewardship

All content needs both an initial assessment and periodic review to assure its scholarly value to the institution, and its suitability to fulfill the CUL's scholarly mission. Additionally, digital content requires frequent re-housing in new architectures, and re-evaluation for long term accessibility. Mitigating threats to digital content often require balancing pros and cons of different technological strategy, gains vs loss. Intellectual stewards inform that process through their expertise as representatives of the scholarly community and the intellectual value of the content itself, and can function as an able co-pilot when making important choices related to migration, discovery and display, and end-of-life assessment. Decisions evaluating the content's value should not be left to the technical support team.

- ☐ Reach out to selectors, curators, archivists and/or other specialists as appropriate for your content to connect their skills and knowledge to the content in your repository.
- ☐ Create and assign role(s) of intellectual steward/curator for each collecting areas represented in the repository. Assign these leveraging the expertise of selectors, and in a way that is resistant to institutional change (obtain buy-in that role conveys with position).
- ☐ Maintain relationships between and repository manager and intellectual stewards/curators.
    - ☐ Develop contact lists (elists, group addresses) for communication about changes in repositories, and use them.
    - ☐ Reach out to intellectual stewards/curators at key times (migrations, upgrades, systems failures, etc.) and solicit advice and help in managing community expectations.
    - ☐ Partnering with intellectual stewards, consider whether each collection is worth preserving. Digital preservation is a significantly costly process, and not all content will likely be worth this effort.
- ☐ Maintain these relationships with regular communication. Include regular check-ins about collecting scope.
- ☐ Draw upon these relationships at key times where users communities are impacted - during interface changes, content migrations, technological end-of-life (both systems and formats), system downtime and upgrades, etc.

Digitizing content

If the content already exists in digital format, then your main concern will be evaluating the quality of the digital objects against what is needed for the chosen system.

If the content is not yet digital, then you will need a plan to digitize it. If your unit has in-house digitization capabilities, you will need to prepare a digitization plan. If your unit lacks such facilities, then you should reach out to DCAPS to discuss how best to digitize the content, and potentially get an estimate from them.

- ☐ Determine how much of your targeted content exists in a non-digital format.
- ☐ Compare your digitization needs to the digitization capabilities of your library or unit.
- ☐ Review current digitization best practices for your content.
- ☐ Develop digitization workflows. Contact DCAPS for advice as needed.
- ☐ If the needs exceed in-house capabilities, contact DCAPS to discuss digitization options.

Analog material and digital surrogates

Digitizing analog content helps expand access and may support preservation of the analog material. However, it should be clear to users of the digital content A.) that it is digitized from a non-digital original and B.) how they can find the original analog material.

Additionally, if the analog content is discoverable in other managed discovery environments, there should be some documentation describing that it has been digitized and how to access the digital surrogates. This two-way relationship between analog and digitized content supports preservation of your digital material.

- ☐ Identify system(s) where analog content is already discoverable.
- ☐ Determine what metadata is needed to define relationships between analog content and digital surrogates. Complete Metadata section of this manual.
- ☐ Consider how evolution of discovery systems and records may affect implicit or explicit relationships of analog content and their digital surrogates over time.

Metadata

Once you have identified the content that will be included in your repository, it is vital to determine whether there is any extant descriptive metadata associated with that content. Depending on what you are working with, you may have access to extensive descriptive metadata or virtually none; it may be in a format that can be easily utilized for the repository or require extensive coding or conversion; and even extensive metadata may not include the fields needed to fulfill the technical requirements of the repository.

Depending on the specific needs of the repository and/or your unit, you may not be personally responsible for making decisions about descriptive metadata, let alone implementing them. But as a repository manager, you are responsible for making sure those decisions have been made.

- ☐ Identified the extant descriptive, administrative, structural, and/or technical metadata associated with the content for your repository.
- ☐ Completed the Metadata section of this manual.
- ☐ Set up an appointment with Metadata Services to discuss extant metadata, and the metadata needs of the repository.

Identifying content creators and rights holders

Digitizing content will often require an investigation of who holds copyright to that content; adding such content to a repository always requires such an investigation. Making this determination is a key prerequisite to any repository process. If you don't know who owns the rights to the content, you should consult with Cornell Copyright Services to help make that determination. Once you have identified the rights holders, it is vital to obtain their permission before moving forward with the digitization and dissemination of the content; Cornell Copyright Services can work with you on this process, as well.

- ☐ Complete the Rights Management section of this manual.
- ☐ Contact Cornell Copyright Services for training on how to identify rights holders and obtain permissions.
- ☐ Identify rights holders for repository content.

Policy and Documentation

Authors: Dianne Dietrich and Erin Faulder

Introduction

The following section details advice for managing your documentation over time. Documentation is an ongoing practice of maintaining your policies and procedures.

Care and feeding of your documentation

Documentation requires upkeep to keep it current and useful. This maintenance requires consideration when setting up documentation in form and function.

All documentation should include a definition of:

- ☐ Stakeholder(s) affected
- ☐ Audience for the documentation
- ☐ Date it was approved

A documentation review schedule should record:

- ☐ Date or frequency of next review date
- ☐ Who or what role should review the documents
- ☐ Who is tasked to make sure the schedule is maintained
- ☐ Date of last revision

The location of documentation should:

- ☐ Be accessible by team that has to review
- ☐ Be accessible by stakeholder(s) and audience(s) for reference as appropriate
- ☐ Provide for a way to easily link related documentation together

Documentation should be preservation-friendly and sustainable. This means ensuring:

- ☐ It is not stored on a single computer
- ☐ There is a way to export from a system if it lives in such an environment (e.g. Confluence)
- ☐ Final versions are saved in preservation-friendly file formats
- ☐ There is a plan to manage revisions and appropriately retire old versions

Policy Writing

Having policies in writing and accessible is a key aspect of any sort of library service, and repositories are no exception. Such explicit policies facilitate making decisions consistently, as well as setting patron expectations. In cases of dispute, having a written policy in advance can help resolve matters in a way which feels fair to the parties involved.

While any policies you set must be consistent with those of Cornell University and the Cornell University Library, there is no one-size-fits all set that can be implemented out of the box. The

policies of a repository must reflect the goals of the sponsoring unit and the intended audience, as well as operating within the larger organizational framework. Some policies may be written for an external audience while other policies or procedures are better suited for internal staff of the repository.

Below, you'll find some common policy areas, and the issues that should be addressed within them. This list is not intended to be exhaustive, but at the very least should prove a good starting point.

Access

These policies govern who may utilize the content within the repository, and under what (if any) restrictions. (Note, this is distinct from accessibility, covered below.) More discussion of these issues can be found in the **Access: Discovery and Delivery** portion of this handbook.

Accessibility

These policies govern the removal of barriers, intentional or otherwise, that stand between valid users and the content within the repository. (Note, this is distinct from access, covered above.) Note: Repository policy exists within a larger framework of obligations, and need to reflect that framework! More discussion of these issues can be found in the **Access: Discovery and Delivery** portion of this handbook.

Collection development and content deposit

These policies govern what material will be included in the repository, how that material is collected, and the means by which it is added to the repository. More discussion of these issues can be found in the **Defining Repository Scope** portion of this handbook.

Preservation

These policies govern the long-term preservation of digital objects contained within and/or delivered by the repository service. More discussion of these issues can be found in the **Preservation** portion of this handbook.

Privacy

These policies protect the privacy of all users of repository services. By default, all repositories abide by the privacy and confidentiality policies of the Cornell University Library; you may wish to consider policies that provide protections above and beyond those.

Student works

These policies govern the inclusion of student works in the repository. While some of these issues are covered in "Collection development and content deposit" above, student works raise special issues, especially around compliance with FERPA. Please see the Cornell University Registrar's FERPA page and Cornell Academic Technologies' FERPA and Technology page for additional information, as well as contacting the Copyright Information Center.

Sunsetting content

This policy governs the decision-making process on whether to retain or remove content. This is one of the sparsest areas of policy, as the assumption at CUL tends to be that once something is added to a repository, it will remain there in perpetuity. But digital collections should be weeded for many of the same reasons that physical collections are, and as such this policy is necessary. If the decision is to retain content, see section on "Exit Planning" in the **Service Planning** section, which deals with migrating content to a new platform.

Takedown requests

These policies govern the response to requests to remove material from repositories, or to redact content from those materials. Contacting the Copyright Information Center when crafting these policies, and when responding to specific requests, is advisable. More discussion of these issues can be found in the **Rights Management** portion of this manual.

Compose, approve and document:

- ☐ access policy
- ☐ accessibility policy
- ☐ collection development and content deposit policies.
- ☐ preservation policy
- ☐ privacy policy
- ☐ student works policy
- ☐ content sunsetting policy
- ☐ takedown request policy

What to document?

The spreadsheet associated with this section – and available as a supplemental file to this document – identifies key decisions discussed in the totality of the Repository Principles and Strategies guide that require documentation in either your policies or procedures.

It has two sections for each decision:

1. Recommended audiences for the decisions by role
2. Possible policies where the documentation could reside

Infrastructure and Interoperability

Authors: Chris Manly and Michelle Paolillo

Introduction

The infrastructure (servers, software, and practices) that support a repository are the unseen foundation of the repository service; this foundation is the "building" in which the content is housed. When everything is going well, nobody is aware of the infrastructure, people visit the house and interact with the contents to get what they need. When something goes wrong, everyone is painfully aware of it, and interaction with the content can be hampered or lost. Repository managers are not expected to be technically accomplished, but do need to have a basic working knowledge of the architectural design of their repository, so that they can choose an architecture that best fits the needs of their uses, and assist with "home repairs" at times when they are needed. Partnering with IT, services providers and/or vendors is inevitable, especially as the service is affected by tool selection, architectural and cost constraints, and unanticipated events. The grasp of the repository manager of the basic challenges and constraints of architectures and technical solutions will directly impact success in communicating with the IT support provider who maintains the system.

Repository managers are encouraged to leverage the resources of the Repository Executive Committee to facilitate the partnerships to work effectively through the checklists below.

The repository manager is the appropriate representative of the service to the user community, whether that be in times of service stability (regarding features and usability), service change (regarding migrations and new features) or service events (like unexpected outages). The repository manager will need to be able to communicate these events effectively to the user community so that technical staff can maintain focus on the support of the technological system(s) in play.

This document can help repository managers as they contemplate designing a new repository, or in response to significant changes in the repository they run. Events such as these may trigger review:

- Recurrent or catastrophic failures of components that support the repository service.
- A large architectural overhaul of a technological component of the repository
- Commercial changes affecting repository components, such as a major acquisition
- Staffing changes inside your repository team; loss of a team member with specialized, necessary skills.
- Loss of funding to support staffing levels, contracts, or systems.

Note that in the case of extreme changes in the categories above, the repository manager may want to contemplate leveraging their exit plan.

Infrastructure

There are a range of options for in-sourcing and out-sourcing of repository infrastructure. Each option on the continuum offers a different level of control, and a different type of direct responsibility.

- An in-sourced solution allows the greatest amount of local control and opportunity for customization of the functionality and for the security of assets and metadata housed in the repository, but it also requires the greatest commitment of staffing resources. There is also the risk of developing technical debt if sufficient staffing isn't allocated to maintaining the infrastructure.
- When considering an out-sourced solution, the repository manager must ensure issues are addressed at a contractual level as there is less direct control over the infrastructure environment. A fully out-sourced solution minimizes the burden of maintenance, but adds the risks associated with loss of control. Outsourced solutions face threats associated with acquisitions, which include loss of control of features and interface design, and even the loss of control for various uses of the content and metadata in the repository.
- Infrastructure is often a combination of in-sourced and out-sourced components. Mixed models of support are acceptable, common, and may meet your needs best.

Definition of service

Complete the **Service Planning** and **Defining Repository Scope** sections of this handbook to assure that you have a scope for your repository. A review of your repository policies may also be helpful, since the architecture should support them. Then consider the additional topics below.

Features

Properly scoped features can support the needs of the repository's users, and contain costs through the prevention of scope creep.

- ☐ Undertake a systematic analysis of user needs for features for the anticipated content of your repository.
- ☐ Rank the features in terms of their priority (must have immediately, nice but can defer for a time, nice but not essential). Explicitly identify what features will NOT be developed (these are out of scope)
- ☐ Use the findings as input to determine appropriate architectures. See "Integrations" at the foot of this document for some special considerations regarding some features that can be provided through other large-scale services.

Cost tolerance

Explore the costs and offerings of the investment of resources, and characterize those resources. Remember that costs need to be reviewed periodically, as financial needs and budgetary support can change.

- ☐ Identify what portions of your infrastructure will be out-sourced
    - Does out-sourced infrastructure exist that meets the needs of your user community of focus?
    - Is funding available for outsourcing? How much? Consider staff effort for
        - system design
        - system implementation
        - ongoing support
    - Outsourcing requires a point person for vendor contact, including escalating bugs that are discovered, service issues, and desires for new features. Determine who will serve in this role.
- ☐ Identify what portions of your infrastructure should be in-sourced
    - Does in-sourced infrastructure exist that meets the needs of your user community of focus?
    - Is staff effort available for the support of these components? How much? Consider system implementation and ongoing costs for support services.
    - Coordination will be required across departments within Cornell. Determine who will perform this function.
- ☐ If your infrastructure utilizes in-sourced and out-sourced components, draw a diagram to express how the components fit together, the flow of data, and area of responsibility for support and funding. This can bring clarity to missing pieces of the overall infrastructure.

Level of service

What level of expectation will you promise to meet for your end users? This affects decisions regarding Monitoring/Availability/Performance.

- ☐ Determine what kind of alerts will you require to alert you or IT minders to the health of the system
- ☐ Make a threat matrix that describes what are the threats to your services, user tolerance for the loss of those services, the likelihood of loss, and likely time required to restore them to better inform your plans for restoration of services.
- ☐ Explore the support plan for each component of your service and determine if the restoration plan aligns with the information in the matrix created above.

Implementation decisions

In this context, "Implementation" is intended to describe the choices associated with assembling the infrastructure components to support the repository service, not the act of system implementation itself.

Service continuity planning lays the foundation of keeping the service running according to acceptable parameters of availability.

- ☐ Use the threat matrix created above to plan for outages and events
- ☐ Document plans and share with appropriate stakeholders
- ☐ Review and update plans periodically

Lifecycle planning devises plans for keeping the architecture up to date, in good condition, and reasonably secure. It keeps on top of developments that can be beneficial for user communities.

- ☐ Monitor development roadmaps for suitable architectures
- ☐ Participate in user communities, forums, and membership calls.
- ☐ Periodically evaluate responsiveness of support for infrastructure components; adjust if it lacks alignment with identified repository needs.
- ☐ Develop an Exit plan; see "Exit planning" in the **Service Planning** section of this handbook.

Change management creates templates for how to introduce service changes to users.

- ☐ Think through possible untoward effects of upgrades and system changes, and make plans for rollback, and identify appropriate points for rollback decisions.
- ☐ Document change plans
- ☐ Consider implementing a formal change management process.
- ☐ Develop testing protocol for the repository components in use. Testing is ideally automated, but can also be manual. If manual, use of standardized scripts is encouraged.

Operations support makes sure the effort to support the system and its service is aligned with its needs.

- ☐ Expect problems identified through service continuity planning and staff for them.
- ☐ Evaluate stack requirements carefully (examples: linux distro, hydra stack, java requirements)
- ☐ Ensure that the components fit within the landscape of the mainstream of supported systems
- ☐ Ensure that staff possess the proper technical skills/get relevant training
- ☐ Ensure that components work well together

Technical constraints

These are not standalone items, but integrated into all components of the infrastructure. They should be reviewed periodically to ensure that the institutional needs are being met.

Backups

Backups of a repository can protect against a few different types of risk. The primary situations to cover are disaster recovery and undesired content changes (either due to accidental changes, or malicious activity by outside entities). Sometimes a single backup solution will cover both of those scenarios, and sometimes a different approach is needed to protect against different risks.

Either way, it is critical to backup both data and metadata in ways that they can be restored coherently with respect to each other. The application components also will need to be either restored from backup or reliably rebuilt in a disaster scenario. Backups are not preservation.

- ☐ Review backup strategy periodically to assure it aligns with needs.
- ☐ Run periodic tests of backup to assure data can be appropriately recovered within the time specified in recovery plans.

Security

There was a time when security could be considered secondary priority for library resources that were designed to be made available to the public. However, due to the change in the security threat landscape, that is no longer the case. Those responsible for repository services need to attend to basic security principles to protect:

- **Data Integrity**: Steps must be taken to ensure that the content hosted by the institution has not been altered.
- **Reputation**: If a system hosted by the institution is compromised and is used as a launch point for further attacks or for visibly malicious activity, the institutional reputation suffers.
- **Protection of Intellectual property/subscription services**: Where copyrighted materials are hosted with restricted access, those restrictions must be protected

- ☐ Identify your security liaison.
- ☐ Keep systems up to date with patches and appropriate configurations
- ☐ Request a regular security review that covers all components of repository infrastructure
    - ☐ Ask the Library Security Liaison for help.

Integration

A single repository architecture rarely provides for all the needs of a specific user community or set of assets. For this reason, integration with other technical architectures may be desired. Considerations for several commonly found integration areas are listed below.

Discovery

Discovery can be accomplished by indexing repository contents and then searching in a local interface, but this has the effect of silo-ing the assets within the repository alone. There are a couple of common methods to open up the silo and make the resources discoverable from outside the repository:

- Allow for crawling by indexers (ex: Google)
- Make index of resources available to other discovery layers.
- Consider the use of a handles to allow for stable URLs for resources. (Can assist in exit strategies should that become necessary.)

Extract, Transform, Load (ETL)

These processes coordinate lifecycle management and complex workflows where repositories and metadata stores are linked serially or recursively. For instance, metadata may be created in one system, but must be populated to another as the system of record (ex: Voyager migration to Archivespace) Likewise metadata from one system might be used to augment a second system. (ex: Scholars and Symplectic Elements). ETL differs from migration in that it is periodic, and keeps live systems in synchrony. ETL processes can affect data and metadata alike; preservation systems often employ periodic, automated extracts of both data and metadata from live repository sites for preservation purposes.

Authentication/Authorization

A repository often has the ability to authorize access based on user affiliation, or to grant/limit access by community affiliation. This can be accomplished by local accounts on the system, but it usually makes sense to integrate with other campus wide identity provisioning systems, especially if affiliations in that system can be re-purposed in the repository. Significant economies of scale in terms of administration support through this type of integration. ex: Shibboleth

Metadata Design and Best Practices

Authors: Jason Kovari and Wendy Kozlowski

Introduction

Metadata is "structured information associated with an object for purposes of discovery, description, use, management, and preservation" (NISO, http://framework.niso.org/24.html). Good metadata facilitates identification and location of content, allows for coordination of instances of an item and supports long-term management and preservation of content.

Select examples of actions facilitated by metadata include:

- Discovering a copy of an article applicable to a class project
- Discover a book in Blacklight, the digital representation in HathiTrust and an image of illustration from the book in SharedShelf or Digital Collections Portal
- Discover a resource and then locate other resources from the same collection
- Uniquely identifying a resource and differentiating between resources
- Monitor fixity of digital resources to ensure
- Build a user experience leveraging logical order among resources or grouping images of the same resource

When trying to understand how best to describe items in a repository, it is important to take into consideration user expectations and needs as well as community standards and best practices; further, it is important to not allow metadata decisions to be determined based on system limitations as metadata inevitably outlives the current discovery environment.

Acknowledging that quality metadata requires involvement from multiple stakeholders, and understanding the intersection of differing priorities, allows for the creation of user focused metadata that is interoperable and reusable across collections both within and beyond the platform in which the metadata were originally created.

Metadata Services (within Library Technical Services) defines and provides guidance for metadata practice across CUL to ensure that metadata is interoperable and reusable across CUL's discovery environments and beyond. This document is intended to help Repository and Collection managers consider the importance of proper metadata design and handling to overall functionality of your repository system. We list questions to consider throughout the process, but the intent is for your repository to work together with Metadata Services to design the most productive yet interoperable repository possible within the repository ecosystem at CUL. You can reach the metadata services group by emailing metadata_info@cornell.edu, and additional introductory information can be found on the metadata services website.

Because decisions and workflows around metadata practice are idiosyncratic and based on individual circumstances, case by case evaluations will be necessary. This document introduces

repository managers and designers to high-level topics for consideration before consulting Metadata Services.

Metadata categories

Generally, metadata can be categorized into a few areas:

1. **Descriptive metadata** enables discovery, identification and retrieval of materials hosted in the repository. It is often the first and most commonly considered metadata type when implementing a repository, and includes elements such as title, author, subject, dates etc.
2. **Preservation metadata** captures details about the resource that is important for long-term retention and management. It may include information about the technical environment, custody and change history, processing, storage and status. It is important that preservation metadata is compatible with the collections management workflow of the archiving institution.
3. **Technical metadata** describes the electronic nature of the resources, such as information about format, file size, checksums, sampling frequencies, how the metadata themselves are collected, and other similar characteristics.
4. **Structural metadata** relates components of a compound resource and/or bundles related objects into a package by describing the way those components are organized. For example, how chapters or pages of a book are related, or in a collection of pictures about a piece of art, what view or perspective each image is of that work.
5. **Administrative metadata** facilitates the management of a resource. It can include such things as information about when and how an object was created, what processing activities have occurred on the object, the documentation of intellectual property ownership, usage and access restrictions and more.

Identifiers are important across all categories of metadata, intended to: uniquely identify a resource; differentiate between resources; coordinate resources across collections and systems; facilitate preservation; manage relationships between resources and collections; and more.

Metadata design

While implementation practice and community norms vary greatly for each type of metadata and across standards, repository managers and metadata professionals should work carefully to build a metadata model that ensures stewardship of resources through the entire digital lifecycle.

Metadata practice versus application functionality

Metadata decisions should not be guided by limitations imposed by the application or user experience; metadata outlives user interfaces and functionalities of any single repository. While it is understandable to want to identify an easy solution by changing metadata practice, doing so has long-term implications for discovery of the resources.

Metadata Services, a central service provider at CUL, is available to help define metadata models and practice to ensure that the repository meets user requirements meanwhile ensuring adherence to standards. Prior to speaking with Metadata Services, it is helpful to consider the following:

- ☐ Have you outlined functional requirements based on known user needs? If so, have you considered what metadata elements might reasonably be provided by a self-service submission, versus which should be administratively applied?
- ☐ Have you considered a minimum level of metadata elements to be required for deposit (self-submission as well as mediated deposit)?
- ☐ Have you identified other repositories or infrastructures that may need to interact with your repository? If so, does your repository environment facilitate data exchange (e.g.: documented APIs, OAI-PMH, etc.)?

Granularity and impacts on scalability

Before talking with metadata services, it may also be helpful to consider the level at which your users will expect resources to be described at and the level at which you can sustainably provide item description. This "granularity" of detail can be thought of two ways:

1. Granularity could be considered as level-of-description, i.e.: should you describe each item, a related set of materials or the entire collection? For example, a series of photographs could be described as "Photos of Ithaca, NY" or as "Photo 1: The Commons, Photo 2: Ithaca Falls, etc".)
2. Granularity could also be considered as detail-of-describe. For example, you could describe an album as-a-whole or describe individually the songs on the album.

Granularity impacts scalability regarding production capacity as well as how well the metadata interacts outside of the original collection and repository; this must be considered as metadata requirements are defined for a set of materials.

Metadata modelling/Metadata strategy

While many repositories at Cornell have established metadata models pre-packaged with the environment, nearly all can be customized; this customization should adhere to documented models as it will create complexity when sharing data beyond the native repository. Further, as CUL continues to move toward open source repositories, such as Samvera-based solutions, underlying data models become more open and flexible. When defining metadata models, it is important to look across and beyond the library community; stakeholders and domain-based communities of practice can offer use cases as well as solutions to guide modeling for different material-types.

In undertaking this work, it is important to understand the communities being served by the repository; for more information on this, see the Curation principle document.

Metadata Services should be involved in the design of metadata models for repositories at CUL. As part of this process, Metadata Services works with repository managers and identified stakeholders to:

- ☐ Determine specifications required or recommended by the repository platform or infrastructure
- ☐ Examine elements and content standards for use in describing your data
- ☐ Identify key pieces of descriptive information that facilitate documented functional requirements (e.g.: data exchange across repositories, metadata faceting, preservation needs, etc.)
- ☐ Create identifier pattern that can be used across materials and collections within the repository

Interoperability

A primary goal of most repositories is for discovery both within and outside of its home environment; due to this need, repositories must support interoperable metadata models. Interoperability, or facilitation of data exchange across systems, is essential for ensuring discovery across environments (e.g.: repository and aggregators) and facilitating metadata synchronization across systems. Further, interoperable models and standardized practice is important in facilitating discovery across collections even within a single repository. Adhering to defined standards and using adequately featured repository environments will facilitate the requirement for interoperability and substantially facilitate a better user experience.

Metadata production

Balancing user needs with best practices

While difficult to say "No" to a user or depositor, it is important to remember that the user who created the set of materials, or who is complaining the loudest, is not necessarily representative of the user base for a collection. Altering metadata practice based on very specific, individual needs creates inconsistent metadata, and can limit possibilities for the reuse of that metadata beyond the original collection.

To aid in responding to user needs, repository managers should consult with metadata services to:

- ☐ Determine user needs and how this may impact metadata creation.
- ☐ Identify who in the repository management group will respond to user requests.

Guidelines/application profile

As part of metadata production, metadata guidelines should be produced, outlining fields used and how those fields should be populated; this documents aspects of the project such as required fields, repeatable fields, expected values and formatting requirements (e.g., Extended Date Time Format (EDTF)), and more. This information should be documented within a

metadata application profile for management of the repository. Further, clearly articulated guidelines aid the metadata capturer in creating consistent and well-formed metadata.

User-generated vs. library-generated metadata

Repository systems often contain a combination of user-generated and librarian or repository manager created metadata. There are benefits to both workflows, but issues related to maintenance and curation of user generated metadata should be specifically considered. CUL Metadata Services can help mitigate problems and will discuss with you things like:

- How best to present metadata elements to encourage their use and communicate importance
- How to teach standard and/or best practices
- Implications of user failure to submit controlled terms (for example how this might limit faceting and browsing within or beyond local environment)

Metadata maintenance and clean-up

All metadata eventually demands clean-up or enhancement; no metadata should be considered static. Normalizing, updating and enhancing metadata is common as we learn more about our resources, as repositories migrate between systems and as the evolution of data affords different models for description.

Metadata Services routinely performs maintenance and clean-up on existing metadata across repositories; these efforts are performed using manual workflows (item-by-item editing) as well as batch processes using scripts and extract, transform and load (ETL) tooling. Metadata services is available to consult on maintenance issues and might be available to perform these tasks.

Documentation

As with all aspects of repository management, documentation around data models and metadata decisions is an essential component supporting long-term sustainability and interoperability. Clear, complete and transparent documentation is important across the entire spectrum of metadata, including the model, guidelines, workflows and decisions made as part of production efforts.

Metadata services can again help you define and begin to create proper metadata about your metadata, but the following factors should be considered when drawing up good documentation:

- ☐ What organizational structure / naming schemes can be consistently used to facilitate creation and finding of the documentation?
- ☐ Where should the documentation be stored so that it is available to all necessary producers/consumers?

- ☐ Who (which specific individual or group of individuals) is responsible for keeping the documentation up to date, and how will they be held accountable?
- ☐ What are the triggers for review and update?
- ☐ What parts of your documentation can or should be made publicly available?

Outreach

Authors: Sarah Kennedy and Jim DelRosso

Introduction

What do we mean by outreach?

Outreach in this context is defined as a programmatic effort to engage in direct contact with the current or desired users and/or stakeholders of the repository in order to advance specific goals (e.g. higher deposit rate).

Who does outreach?

In an ideal world, each repository would have a dedicated outreach and communication strategist to initiate, plan, and implement all outreach efforts associated with the repository. In practice, however, the effort is often much more dispersed, with the Repository Service Owner serving a central coordinating role, assisted by liaison librarians, content selectors, service sponsors, and Library Assessment and Communication as needed.

Reason/Goal

The first step for developing an outreach program, and specific efforts within that program, is to identify why you're doing it. Distill the main goal(s) of your outreach program and efforts such that they are specific and measurable. Clearly stating the goals of your program will make subsequent steps of the process more focused.

Examples may include:

- Increasing deposit rates of born digital faculty publications by X percent;
- Promoting general awareness of the repository service as a resource for self-archiving; or
- Directing users to high-quality, domain specific open access resources.

- ☐ Clearly and succinctly articulate your main outreach goal
- ☐ Identify any specific deliverables
- ☐ Choose appropriate assessment intervals

Audiences

The most effective outreach strategy is one that is crafted for a specific audience. Repositories serve a wide variety of different users and stakeholders, so when considering outreach as a whole, start by brainstorming a list of potential audiences.

Think of all the roles identified in this handbook, as well as other stakeholders your repository work to date has surfaced, and other contributors or user groups. If you've identified similar projects within your institution (or beyond), see what audiences they have identified, and investigate whether you have parallel audience groups. Outreach audiences may exist within your organization, within your institution, or beyond the walls of both.

When working with audiences outside of CUL, please remember that Cornell University Library Assessment and Communication can provide assistance with any of the following steps as you work toward creating your own outreach plan.

- ☐ Brainstorm a list of potential audiences
- ☐ From that list, identify specific intended audiences
- ☐ If you've identified external audiences, reach out to Library Assessment and Communication

Potential partners

Once you know your goal and audience, identify the partners who will provide the best support to your outreach effort. Consider their proximity to relevant user groups, familiarity with collections, and proficiency with required communications. Note that your outreach partner for one effort may be the audience for another, and vice versa.

Some common outreach partners include:

- Service sponsors: Often serving at the senior management level (e.g. AUL or Director), the Service Sponsor may champion the repository service in interactions with other campus leaders at the level of Department Chair or Dean, or represent specific outreach tasks that require oversight from senior management (e.g. large scale permissions analysis projects). The Service Sponsor may also be responsible for representing outreach efforts at an internal level, acquiring sign-off on new outreach efforts from the University Librarian.
- Content selectors: With their in-depth knowledge of the collections, content selectors may provide valuable insight into the audiences and stakeholders who would be most appropriate for outreach around a given set of material. They may be particularly helpful in communicating the scholarly or practical value of a particular collection, or in identifying venues and delivery mechanisms for sharing announcements.
- Liaison librarians: Liaison librarians can communicate the value of a repository service to a campus audience. The Repository Service Manager should ensure that liaisons feel prepared and confident delivering a message on behalf of the outreach team. A train-the-trainer session may be helpful to ensure that all parties are comfortable with the messaging.
- CUL Assessment and Communication: Our colleagues in CUL Assessment and Communication have expertise in marketing and branding, as well as access to a number of exclusive communication channels for reaching both internal and external audiences. Contact libcomm@cornell.edu for assistance.
- Campus partners: There are a number of academic units or service centers on campus (e.g. the Center for Teaching Innovation) who might be able to provide support on a given outreach effort, depending on their function. Be sure to consider who else on campus might be trying to achieve similar outreach goals, even if they're not part of the library.

- ☐ Identify potential outreach partners and define their roles, including external partners
- ☐ Identify training needs that outreach partners may require
- ☐ Implement training programs as needed

Prioritization

When you have identified who your audiences are, and why you would reach out to them, you can begin making decisions on work prioritization. This is a decision that will likely rest with the Repository Service Owner, but also represents an opportunity to connect further with internal stakeholders like the Service Sponsor or liaison librarians. Keep in mind that outreach performed for the repository can often be leveraged to serve other institutional goals.

These stakeholders can also help you surface external factors that might impact the success of your efforts: other outreach efforts targeting a given audience, issues arising from the academic calendar, etc. The overriding goal of this stage is to determine which audience(s) to address first, and when.

- ☐ Match goal(s) of outreach to each audience.
- ☐ Prioritize the audiences/goals to determine your list of outreach efforts.

Value statement (your pitch)

With an audience and goal in mind, craft your message around a convincing value statement, or pitch. This is where you deliver the "so what?" of your message in a way that addresses the immediate need or pain point of the audience. A strong value statement considers audience pain points and addresses them directly, avoiding jargon that could be confusing. It succinctly explains how your service adds value and includes a call to action when appropriate.

Consider the language, tone, and length of your outreach message. If you are asking something of your audience, consider whether your word choice conveys a tone that the audience *must* or *should* or *might consider* performing a particular action. Craft your pitch with a mind to your audience's level of expertise (which might exceed your own).

Be mindful of word choice and generally aware of how the message will come across to a diverse audience of readers. Your word choice may convey a variety of subtle messages (e.g. urgency, excitement, authority, friendliness/ approachability, etc.) and/or reflect implicit biases that you hold about your audience(s). Before you send the final draft, be sure to solicit feedback from one or more volunteer test readers who are themselves representative of the diversity of your audience(s).

Typically, you won't have much time or space to make your point, sometimes due to constraints of the delivery mechanism (e.g. character limits on social media). Be as succinct as possible, always providing the audience the opportunity to seek additional information if desired.

- ☐ Craft one or more value statements that distill the most important benefits of your repository service to your particular audience
- ☐ Scan all language for unnecessary jargon or terminology with which your audience may not be familiar
- ☐ Consider word choice, especially if you are requesting your audience to do something
- ☐ Solicit feedback from diverse test readers
- ☐ Proofread final outreach message, omitting needless words

Delivery mechanisms

Channels for reaching internal audiences may include communication from senior library leadership, the Cornell Chronicle for news items, the Library Update, the CUL home page, digital signage, flyers and posters, CUL branded brochures, email lists, online newsletters, and Ezra Updates. Channels for reaching external audiences may include various media outlets and listservs. CUL Assessment and Communication can also assist with crafting surveys and other tools for gathering user data.

While email has, for better or for worse, become the vehicle de rigueur for many outreach efforts, it is by no means the only available vehicle. Others may include:

- Social media
- Video or screencast
- Website
- LibGuide
- Newsletter
- One-pager (digital and/or print)
- Workshop
- Digital signage
- Press releases
- Branded collateral (flyers, magnets, pins, pens, stickers, general swag, etc.)
- 1:1 Research Consultation
- Departmental e-lists

Please note that the use of many of these delivery mechanisms are governed by policies set at the unit, CUL, or University level; make sure to familiarize yourself with those policies before sending your communication.

- ☐ Select the most appropriate delivery mechanism (s) for your audience, goal, length of message, timeline, and available resources
- ☐ Familiarize yourself with policies governing delivery mechanism before sending your communication, consulting with CUL Assessment & Communication as needed
- ☐ Identify which partner is responsible for delivery

Preservation

Authors: Michelle Paolillo and Erin Faulder

Introduction

This section may help familiarize the terms digital preservation and digital curation. Together, digital preservation and digital curation are future-focused efforts that encompass the digital lifecycle within a repository framework.

It is important to define digital **preservation** for the purposes of this document, since the term is used variously. "Digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change" (Wikipedia, https://en.wikipedia.org/wiki/Digital_preservation). Digital preservation differs from backup, which although useful, does not anticipate problems of media deterioration or obsolescence, or file format obsolescence, nor do they introduce strategies to detect and protect against spurious or accidental unwanted changes. Nor do backup copies or mirror sites ensure adequate description of content (including technical and administrative metadata) in the way that preservation policies do. Finally, backups do not entail procedural planning over the long-term, as digital preservation does. Repository managers who are unfamiliar with the subject of digital preservation may be well served by the tutorial Digital Preservation Management: Implementing Short-term strategies for Long-term Problems.

In addition to digital preservation, digital **curation** is another helpful concept. "Digital curation is the selection, preservation, maintenance, collection and archiving of digital assets. Digital curation establishes, maintains and adds value to repositories of digital data for present and future use" (Wikipedia, https://en.wikipedia.org/wiki/Digital_curation). Our physical collections benefit from the curatorial efforts of selectors, archivists, and other specialists who are well acquainted with the scholarly significance of the collection. Likewise, decisions regarding the content of digital repositories should benefit from curatorial efforts, especially at key decision points that could affect the way material is presented digitally (migration to a new platform, user interface changes, the addition of new features, de-duplication, etc.).

Getting started

Digital preservation and digital curation are focused on content and keeping content available over time, not necessarily on the technical form that houses the content. Digital preservation relies in part on actions that occur before content is in a preservation environment, so understanding how you are collecting content will inform what is necessary and practical to expect from these pre-ingest environments. When thinking about preservation and curation,

think about planning for 15+ years down the road, not 3-5. Consider what could happen in that span of time: platform migrations, catastrophic system failures, changes in access of content, retirement of key personnel with institutional memory, inter-institutional efforts, changes in institutional support, evolution in collecting focus. Digital preservation also includes preserving and maintaining documentation that articulates standards used over time and the arrangement of content itself. The documentation may or may not be stored externally to the system housing the content.

If you need help with the review and preservation and curation planning for the content of your repository, please contact cul-digpres@cornell.edu for assistance.

Mitigating external threats

Do you have a strategy in place to mitigate external threats to the content in your repository?

Preservation over time includes planning for risks that may threaten the health of your repository. These threats may include inadequate staffing, loss of institutional knowledge that is undocumented, economic threats (budget cuts), and system failures.

- ☐ Identify and document potential external threats to your repository.
- ☐ Evaluate likelihood of threats and extent of risk.
- ☐ Develop a plan to mitigate the risks.
- ☐ Review and evaluate architectural design of your repository for preservation capability (see infrastructure principles). This will inform basic strategy for preservation.
- ☐ Review institutional preservation commitment to assets and metadata. This will inform as to appropriate level of resources assigned for preservation.

Content integrity and authenticity

Are you ensuring integrity and authenticity of digital content?

Preserving digital content requires a level of trust that users discovering the content access objects that are authentic and have integrity. Authentic objects are those that are what they claim to be (e.g. a dissertation is actually Jane Coriander's dissertation). Evidence of authenticity is documentation around who submitted or ingested the content and verification that metadata describes the right object ingested. Integrity means the object accessed is the same as the object when it was put into the repository. Evidence of integrity is often a checksum created when the object is ingested that can be verified on access.

- ☐ Create checksums for all content ingested into repository and store the value.
- ☐ Recompute and verify checksums against a stored value on a scheduled basis to look for changes to the content.
- ☐ Consider allowing users accessing content to also access the checksums to verify integrity of object.
- ☐ Review content collected passively to ensure object submitted is what submitter described and intended to submit.

Obsolescence planning

Do you have a strategic way to mitigate risk of file format obsolescence?

If not, develop key policy and enforce by technical means if possible.

- ☐ Develop a policy on accepted or recommended file formats that have greater likelihood of enduring into the future.
    - Note that the collection style of your repository and the actual content you hope to collect may dictate the level risk that you deem "acceptable". In general, active collection of content may allow for more control over formats. In tension with this, some content may only be available in riskier formats.
    - Example: eCommons has a policy of recommended file formats that it makes available to depositors. Note that the chart is subdivided into basic content types, and gives a basic risk level for each format type. This chart is periodically reviewed for the inclusion of new formats.
- ☐ Normalize deposited files to a standard format likely to be persistent.
    - Example: arXiv employs the strategy above to accept a constrained list of file formats, and in addition converts the articles submitted to PDF's. Even though the original submissions are of a variety of formats, this normalized version, the PDF, is a uniform derivative with a high probability of longevity. arXiv mitigates risk of obsolescence, by keeping both the original formats and the normalized derivative.

Minimum metadata requirements

Do you have minimum metadata necessary to support long-term preservation management?

Establish a set of minimum metadata required for deposited content, and ensure that all deposited content meets this standard. "Preservation metadata" can be understood as an amalgam of other types of metadata: descriptive, administrative, structural and technical. These metadata types can be kept in one schema, such as PREMIS, or they can be kept separately in various components of the repository system. Consistent and well-documented metadata supports preservation curation activities over the long-term (15+ years).

- ☐ Review and document where metadata is stored in repository. (Partner with Metadata Services to ensure uniform and thorough review.)
- ☐ Establish policy for minimum metadata on deposit. (Partner with Metadata Services to ensure completeness.)
- ☐ Use standard, open metadata schemas where possible.
- ☐ Ensure that all schemas in use are fully and externally documented, in an open, preservable format.
- ☐ Determine which constrained vocabulary(ies) are applicable to the content being collected
- ☐ Determine if metadata minimums and standards are enforceable technologically.
- ☐ Maintain documentation of changes to metadata minimums and standards over time. Use open and preservable formats.

Essential documentation: Rights

Do you have documented rights in place to facilitate legal obligations of creating multiple preservation copies and format changes over time?

Digital preservation always requires creating multiple copies and may may require format normalization over time. Intellectual property law and specialized donor agreements makes the legality of file format migration and creating multiple copies for preservation murky unless permission is explicitly given. Ensure that your rights management strategy ensures that the curator of the content will have the right to appropriately preserve the digital content.

Contact copyright@cornell.edu to obtain guidance.

- ☐ Create a standard agreement for content depositors to transfer rights to the content that allow curator or repository staff to perform preservation activities.
- ☐ Decide who is on the hook for vetting or defining rights that are transferred.

Essential documentation: Accessibility and review

Is essential documentation of your repository appropriately available and accessible?

All forms of documentation should exist outside your repository, and the availability of documentation affects a wide variety of stakeholders. This strategy will allow the referencing of important information by developers in the case of failure of the repository system. It can also allow for more effective participation of users when submitting content, and can assist in managing expectations about collecting scope and delivery mechanisms.

- ☐ Store these types of documentation outside the repository
    - Manifests for repository content covering both data and metadata
    - Schema and data arrangement
    - Defined roles and representatives, decision making processes, technical infrastructure, policies
- ☐ Review all documentation periodically and keep it updated
- ☐ Clearly record where documentation is kept, and record access credentials where needed for future repository managers.
- ☐ Review where documentation is kept over time with an eye toward resilience over time.
- ☐ Turn all preservation decisions and strategies into a plain-language document for curators, content creators, and users to clearly establish expectations.
- ☐ Refer to the **Policy and Documentation** section of this handbook

Rights Management

Authors: Amy Dygert and Jim DelRosso

Introduction

Rights management concerns what the repository can do with content submitted for deposit. It involves a basic understanding of copyright law as well as an analysis of the creation and ownership of content for deposit.

N.B. Some of the questions below will need to be answered for each new collection, and sometimes for each new piece of content. Many repositories contain a range of content with a range of different rights situations, so you should be prepared to engage with these questions frequently and as needed.

These principles concern deposit by repository managers. When repositories permit users to self-deposit, they must also employ a mechanism to ensure that users have the authority to deposit the relative work(s). Requiring users to affirmatively select and authorize licenses for their work places rights management responsibilities on users. Examples of deposit agreements and licenses can be found at arXiv and eCommons.

Who has the Authority to Deposit Work in a Repository?

Only copyright owners have the authority to deposit a copyrighted work into a repository. A content creator is typically the first copyright owner. As noted above, if the content creator has licensed or transferred his/her copyright, s/he most likely no longer has authority to permit deposit of the work into a repository, depending on the terms of any license or transfer agreements.

Because the copyright owner may so easily change, repository managers will need to evaluate whether they have authority to deposit a work based on a number of facts that help establish copyright status. Among these facts are copies of any agreements that govern ownership and use of the content.

- ☐ Identify copyright owners for all content to be added to the repository.

How do I Perform a Copyright Analysis?

Generally speaking, content was, is, or is not protected by copyright. You should assume content is protected by copyright, although there are a variety of reasons why the content nonetheless may be used. It may have been protected by copyright at one time, but now be in the public domain. It may not be eligible for copyright protection. It may be protected by copyright, but subject to a legal exception or a private license or agreement between the author or publisher and consumers. It may have a Creative Commons license. It may have been licensed through a subscription service. The following sections will assist repository managers in performing a cursory copyright review. For more complicated scenarios, repository managers

should collect the data indicated in the Documentation section and contact the Copyright Information Center at copyright@cornell.edu to further inquire about the status of a work and how it may be used.

1. ***Is the content eligible for copyright protection?***
    a. Is the work an original work of authorship, including a literary, dramatic, musical, artistic, or other intellectual work? Is it in a fixed, tangible form? Does it contain a modicum of creativity?
        i. If yes, continue with analysis.
        ii. If no, the work is not eligible for copyright protection and may be used.
    b. Is the work a trademark, slogan, patent, invention, or trade secret?
        i. If yes, the work most likely does not have copyright protection, but may have another legal protection. Consult with the Copyright Information Center at copyright@cornell.edu.
2. ***Is the work in the public domain?***
    a. Was the work published before 1923?
        i. If yes, the work is in the public domain and you can use it as you'd like.
        ii. If no, continue the analysis.
    b. Was the work published between 1923 and 1977 without a copyright notice?
        i. If yes, the work is in the public domain and you can use it as you'd like.
        ii. If no, continue the analysis.
    c. Was the work published between 1978 and 1989 without a copyright notice, and without registration within 5 years?
        i. If yes, the work is in the public domain and you can use it as you'd like.
        ii. If no, continue the analysis.
    d. Works created between 1923 and 2002 may or may not be in the public domain depending on whether they were published, registered, registry renewed, or contain a copyright notice. To determine the copyright status of a work, consult Copyright Term and the Public Domain in the United States, the Copyright Digital Slider, or contact the Copyright Information Center at copyright@cornell.edu for assistance in making a proper determination.
3. ***Does any permission, special license, contract, or other arrangement apply to the work?***
    a. Did the author give you permission to use the content? Is there a written agreement that describes the permitted use? Is there a deed of gift? What does it allow?
    b. Is the content subject to a special license? What do its terms allow?
    c. Does the content have a Creative Commons license? What type of CC license is it?

☐ Become familiar with process of copyright analysis.
☐ Evaluate content for inclusion in repository.
☐ Contact copyright@cornell.edu for clarifications as needed.

What Documentation is Necessary to Determine Copyright Status?

Repository managers will need as much of the following information as possible to perform an educated copyright analysis. Frequently, not all of this information will be available; but the more of it a repository manager has access to, the more informed their copyright analysis will be. Relevant information to collect includes:

- Creator information: Author's name, date of birth, data of death, obituary, a list of heirs (spouses, children, grandchildren, parents, siblings, nieces/nephews)
- Copyright owner information: Author? Publisher? Employer? Individual and corporate names, including any name changes and dates after creation of the content.
- Date of creation
- Date of publication, if any
- Any evidence of copyright transfer (e.g., publishing agreements, estate transfers)
- Any other agreements or licenses (e.g., deed of gift, Creative Commons license)

In addition to obtaining this information, it's important to keep it in such a way that it can be easily located and referenced later -- and by later, we mean at any point while the object is still being kept by CUL. Where possible, this information also should be retained as metadata. If that cannot be done with the system being used, workflows should be established so that permission documentation can easily be found. In either case, plans should be made to keep the information for as long as the object is kept, including transfers between systems and managers.

- ☐ Obtain sufficient documentation for determining copyright status.
- ☐ Work with Metadata Services to incorporate this documentation into metadata if possible.
    - ☐ Establish alternated means of keeping documentation if it cannot be incorporated into metadata.

What if there is content in my repository that may not have been through these processes?

Such situations will require a risk analysis to determine the best course of action. Contact copyright@cornell.edu to discuss how to proceed.

Research Data

Use and attribution of research data is contemplated by Introduction to Intellectual Property Rights in Data Management published by the Cornell Research Data Management Service Group.

What do you do when someone asks you to remove an item from the repository?

Takedown notices can be worrisome, especially those that come with threats of legal action. However, by following the steps above, you'll be better prepared to deal with them. In general,

CUL and Cornell University both support denying such requests if the copyright and permissions analyses for the item in question were performed correctly. Nonetheless, it's also important to have a policy in place for how to respond to the requests you receive. One key distinction is whether the request comes from the content owner or not. How you respond to requests from the content owner can impact your relationship with that entity, and if they're a faculty member or a donor, that's an important consideration. If the request does not come from a content owner, and your copyright and permissions analyses were correct when the item was uploaded, you likely should not take down the item. Having a policy in place to refer to in communicating with the requester can be very helpful.

N.B. Don't hesitate to contact copyright@cornell.edu or University Counsel if you're feeling unsure or threatened. Some requesters will invoke the possibility of legal action, and there are systems in place at Cornell for handling that.

- ☐ Takedown policy established.
    - ☐ Policy has been discussed and approved by Copyright Office
    - ☐ Policy accounts for requests from content owners.

Understanding Applicable Copyright Law

Six exclusive rights

Copyright laws are a federal laws that give the owner of a copyright exclusive rights to use their work in one of six ways:

1. To reproduce the work
    - e.g., make physical or digital copies of the work for colleagues, students, repositories
2. To prepare derivative works based upon the work.
    - e.g., prepare a subsequent article, chapter, or book that builds upon the original or prior research on a particular topic
3. To distribute copies of the work
    - e.g., distribute physical or digital copies of the work to colleagues, students, at conferences, in repositories
4. To publicly perform the work
    - e.g., show video of field work in the classroom or at conferences
5. To publicly display the work
    - e.g., show photos, exhibits, and figures from works in the classroom or at conferences, post articles in repositories
6. To publicly perform sound recordings via a digital audio transmission
    - e.g., for those working with sound recordings, to digitally transmit works (broadcast online, etc.)

What are exclusive rights?

"Exclusive rights" means that only the copyright owner has the authority to engage in the six aforementioned rights. However, because knowledge and society would fail to progress if only a copyright owner could engage with copyrighted works, there are two ways in which others are legally permitted to use copyrighted works. These are referred to as copyright limitations.

Copyright limitations

Copyright is limited in two ways: statutorily and contractually.

Although there are several statutory limits, the two most commonly known among researchers and librarians are fair use (Section 107) and the libraries exception (Section 108). When a statutory limitation is applied, the copyright owner still owns the work and the copyright in it, but the law permits others to use the work under certain circumstances. When a copyright expires -- currently 70 years after the death of an author, although there are several nuances to this rule -- the work is no longer protected by copyright and the public can use the work in any way they desire.

Contractual limitations arise when a copyright owner engages in a private agreement with another party that affects the ownership or use of the copyrighted work. The six exclusive

rights discussed above are commonly referred to as a bundle of rights because copyright owners control each of the rights individually and as a group. When a copyright owner contracts with another party to permit use of their rights, the owner can give away one, some, none, or all of their rights. He or she can transfer or license the rights. He or she can enter into an exclusive or non-exclusive, irrevocable or revocable license.

Contractual limitations on copyright often occur because an author has signed a publishing agreement that transfers or licenses copyright of the work to the publisher. Many academic authors unwittingly transfer their copyrights to the journals in which they publish. Authors who seek to deposit their works in repositories must refer to any publishing agreements they have to determine whether deposit in a repository, or any other use, distribution, or publication is permitted. Many journals permit deposit of a pre-print version or after an embargo period. Authors and repository managers must always consult the publishing agreement that the author signed before blindly depositing a work that has been previously published elsewhere.

Work Made For Hire

Another contractual limitation on copyright can occur in a work made for hire ("WMFH"). A WMFH is a copyrighted work made by one person but whose copyright is owned by another person or entity. The most common WMFH situation is an employee/employer relationship, although it can also arise when two parties agree that copyrighted work is a WMFH (e.g., when a person is specifically commissioned to create a copyrighted work for another).

At Cornell, copyright ownership of all works of authorship remains with the author, except under certain circumstances. Generally, these circumstances include the creation of work (1) by nonacademic appointees within the scope of their appointment, (2) by an academic appointee pursuant to a specific direction or assigned duty, (3) when developed under a sponsored research or other agreement that confers copyright ownership on the university, or (4) that was developed with substantial use of university resources. See the Cornell University Copyright Policy Statement, Policy 4.15. On page 14, Appendix A: Flow Chart, Creative Work Copyright Ownership Determination is particularly instructive.

Orphan Works

One additional consideration when contemplating copyright ownership is that of orphan works. These are copyrighted materials whose owners cannot be identified or contacted in order to obtain permission for use. Those wishing to make use of copyrighted works should make every effort to find a copyright limitation (e.g., fair use) or copyright owner permission to use the work. Seek permission by searching for the owner, publisher, or heirs of each. For thorough instructions on conducting a copyright owner search, refer to the Society of American Archivists Orphan Works: Statement of Best Practices. The appendices, which chart a course of action for searching for copyright owners, is particularly helpful.

If the copyright owner cannot be found despite a thorough search and best efforts, the work may be deposited in the repository with certain caveats. First, the work must not be commercially available at a reasonable price. Second, proof documenting the search must be retained. Third, the material should be posted with the accompanying statement:

- *After conducting a thorough investigation, [Repository Name] has identified [Item Name] as an orphan work. [Repository Name] would like to hear from individuals or institutions that have any additional information about the rights holders of [Item Name]. Please contact [Contact@Address] to provide more information.*

Finally, should a rights holder be identified and contactable, effort should be made to contact the rights holder and obtain permission for deposit. If a rights holder refuses to grant permission and fair use is not applicable, the work may need to be removed from the repository.

- Review applicable copyright law.
- Keep this information in a place that can easily be referenced.
- Contact copyright@cornell.edu to clarify any questions.

What About Creative Commons and RightsStatement.Org?

Creative Commons and RightsStatements.org are two initiatives to create simple-to-understand terms to help users decipher how they can use a copyrighted work. Creative Commons licenses are applied by content creators who want to communicate to the public whether and how they permit others to use their copyrightable works. Creative Commons offers six different licenses, all of which say that the underlying works are protected by copyright, but that their creators permit public use under certain circumstances. RightsStatements.org created 12 different statements divided into three categories (works that are in copyright, works that have no copyright, and other) with the goal that cultural heritage institutions would classify the digital works in their collections so that patrons would know what they could do with a work.

The two conventions can intersect where an asset in the library's collection has a Creative Commons license, and the library wishes to convey what that license means to patrons. Here, the library would select one of the 12 RightsStatements to provide an additional layer of information as to the copyright classification of the work.

Use of RightsStatements.org and/or Creative Commons is not mandated by Cornell University Library.