POTENTIAL USES

Once you have decided that your organization needs an IR and you have the necessary institutional support and commitment from the right mix of people, you still have many questions to work through before deciding on an actual IR system.

You can equate this exercise with the tremendous amount of research and preparatory work that goes into preparing an RFP (request for proposal). The more informed you are about how your organization will go about using its institutional repository, the better the decisions you will make as to which IR system to use and the organizational structure needed to sustain it.

As discussed above, the suite of services that an IR offers should and must be determined by the unique needs of the institution. As a result, no boilerplate RFP exists that will work for every organization. Each organization must attempt to answer the following questions on its own.

How will the IR be used?

Although predetermining all possible uses of the IR is not necessary, having some idea of how the repository is going to be used is crucial. This knowledge may not only narrow the field of potential systems, but it also could bring to light some potential allies and early supporters in your organization.

You can explore the question of how an IR might be used in several ways. The best method depends on the size and culture of an institution and may be a combination of methods.

If the members of an organization are already familiar with the concept of an IR, then a survey might be the best assessment tool. The survey should inquire about potential existing collections that could be migrated into the IR. Then follow up to gather specific details about those collections to ensure you fully understand their scope and requirements.

For instance, if a potential collection is of datasets for publicly funded grants, then knowing the datasets' size and format types is useful. For what length of time do the datasets need to be made available? Will all the datasets share the same level of access or is there a need for different access levels (such as world-wide, institution, or research lab)? Are there alternative repositories already available?

The focus of another survey should be about the types and scope of individual, collection-less documents. This focus is especially important since many people may not think of their independent documents as being potential candidates for an IR collection.

Through the survey you may discover that your community members have many conference papers, for example, that could be brought together into a single institution-wide collection. Only through the macro-level view provided by the survey would you see the potential for institution-wide collections.

If the organization is unfamiliar with the concept of institutional repositories, then a survey about potential collections and documents would *not* be

helpful, since people would likely not be cognizant of what might constitute appropriate material.

Instead, gather information about the organization's receptiveness to the technology and practice. The survey could include questions about current use of e-print repositories, such as arXiv for physics papers and RePec for research papers in economics.

Other questions to ask include:

- Where do people currently store their digital documents?

- How often do they e-mail or mail documents to colleagues within your organization? At other institutions?

- Do they maintain work-related homepages? If so, what types of materials are there?

- Ask them to rate their level of confidence that their unpublished gray literature will be retrievable in five years? In 10 years?

If all parts of your organization are well-represented, the survey should be able to reveal if and where a need for an institutional repository exists.

In small organizations, surveys can be replaced by one-to-one or one-to-few conversations. At a larger institution, in-depth inventories of IR-appropriate materials from a sample of community members could be extrapolated into estimates of potential institution-wide uses.

Another way to help determine how an IR might be used is to take a look at your organization's current website. Is the website purely informational, or is the website used to distribute locally authored documents? Those collections of locally authored documents distributed *via* your institution's website are the low-hanging fruits for your IR.

As an example, here is a small sampling of the types of documents distributed through the University of Rochester website:

- Rochester Center for Economic Research Working Paper series

- W. Allen Wallis Institute of Political Economy Working Paper series

- University of Rochester Computer Science Technical Reports

- Archives of SimonBusiness, a semiannual publication of the William E. Simon Graduate School of Business

- Collection of publications extending back to 1990 authored by members of the Conversational Interaction and Spoken Dialog Research Group

- Online *Journal of Undergraduate Research*

- *Invisible Culture*, an online journal for visual and cultural studies

This list does not include the several hundred homepages of faculty members that contain links to preprints, conference presentations, datasets, and other forms of digital scholarly material.

Most, if not all, of these collections and documents on the University of Rochester's website could reside within the IR. Moreover, the argument can be made that doing so is beneficial and efficient.

When trying to determine how an institutional repository might potentially be used by your organization, your community should look at the types of collections that already populate or are being considered for established IRs.

The collection types highlighted below do not comprise a comprehensive list. IRs are still in their infancy, and the full suite of potential uses has yet to be uncovered.

Eprints

When Paul Ginsparg created arXiv in 1991 to facilitate the sharing of electronic preprints (e-prints) in high-energy physics, he revealed the tremendous value of e-prints to the discipline.

Today, arXiv averages more than 3,000 e-print deposits per month and has become a well-established member of the scholarly communication paradigm in physics. Following the success of arXiv, e-print servers have been created in numerous other disciplines, with varying success.

Institutional repositories make excellent homes for e-print collections. They could contain collections of e-prints authored only by members of the institution or collections that are open to deposits from around the world, such as arXiv.

Examples of e-print collections in IRs:

- Australian National University's EPrints Repository—http://eprints.anu.edu.au

- CERN's Document Server—http://cds.cern.ch

- Max Planck Society's eDoc Server—http://edoc.mpg.de

- Universität Stuttgart's OPUS—http://elib.uni-stuttgart.de/opus/doku/english/index_english.php

Working papers and reports

With the advent of digital communication, the number of newly established working papers and report series has grown exponentially because most of the production costs, including paper, printing, and distribution, have been eliminated.

Often the working papers and report series live on a website, particularly those of academic departments and institutes on a university campus. These collections are ideal content for an IR and are often the first collections to appear in newly established IRs.

Examples of working papers and report series in IRs:

- European University Institute's Robert Schuman Centre Working Papers—http://cadmus.iue.it/dspace/handle/1814/811

- Iowa State University's Computer Science Technical Reports—http://archives.cs.iastate.edu

- MIT's Sloan School of Management Working Papers—https://dspace.mit.edu/handle/1721.1/1792

- NASA's Marshall Technical Reports Server—http://trs.nis.nasa.gov

- Yale Law School's John M. Olin Center for Studies in Law, Economics, and Public Policy Working Paper Series—http://lsr.nellco.org/yale/lepp

Conference proceedings

For the host of a conference, creating the conference proceedings can be the most difficult, time-consuming, and costly part of the process. Too often, the effort goes unrewarded when the conference proceedings receive only limited use because distribution is minimal.

Most IR systems contain all the necessary functionality to automate the process of creating digital conference proceedings. Responsibility of providing and submitting the appropriate versions of the papers can be pushed down to the level of the authors, and distribution need be little more than the sharing of the collection's persistent URL.

Meanwhile, search engines such as Google and OAI metadata harvesters provide avenues of discovery for countless numbers of potential readers from around the world, including those who never realized such a conference existed.

An excellent example is the *Proceedings of the 4th International Symposium on Cavitation*. The conference was held on the campus of the California Institute of Technology in June 2001.

At the same time that Dr. Christopher Brennen was beginning to organize the conference, the Caltech Library System was starting work on its institutional repository system. Through a collaborative effort between Brennen and the Caltech Library System, the *Proceedings of the 4th International Symposium on Cavitation* were the first in the series to be collocated, published, and distributed electronically.

Brennen believes these proceedings are more widely available than any of the previous ones, which is supported by the fact that the conference proceedings are a top-20 hit on Google with the search term of "cavitation."

Examples of conference proceedings in IRs:

- *Proceedings of the Annual New York Wine Industry Workshops*—http://ispy.library.cornell.edu:8080/handle/1721.1/157

- *Proceedings of the 4th International Symposium on Cavitation*—http://cav2001.library.caltech.edu/

- *13th International Congress of Slavists*—https://dspace.gla.ac.uk/handle/1905/21

ETDs

Collections of electronic theses and dissertations (ETDs) are common content for an institution's repository. With few exceptions, a thesis or dissertation is in a digital format, such as Microsoft Word or LaTex, until the author presses the computer's print button at the end of the process. In most cases, the digital files are then stored by the student on a disk and within a few years the files become inaccessible due to degradation or obsolescence of format.

An IR offers the students an alternative—a place into which they can deposit the digital files that comprise their dissertation, which then are cared for by

the institution and made widely and easily available. Moreover, as discussed in Chapter 3, the availability of ETDs can dramatically increase use of the dissertations and theses.

A collection of ETDs also can benefit faculty. Faculty members often store digital copies of their students' dissertations for their own research purposes, because graduate students, particularly in the sciences, often write dissertations on topics that are tangential to the research of their faculty advisors.

ETDs are an example of content that is ideal for an IR because they have institutional value. Moreover, establishing ETD collections provides real benefits to students and faculty alike.

Examples of ETD collections in IRs:

- Boston College—http://dissertations.bc.edu/

- California Institute of Technology—http://etd.caltech.edu/

- University of Edinburgh—www.thesesalive.ac.uk/index.shtml

- Utrecht University—http://dispute.library.uu.nl

- Virginia Tech—http://scholar.lib.vt.edu/theses

Datasets

Collections of datasets represent a different type of use of an IR. Datasets most often serve as primary source material on which research is based, rather than as the final output of scholarship. But a rapidly growing need exists for a means to facilitate the storage, preservation, and distribution of datasets.

Part of this need is driven by the increasing demand for sharing data generated by publicly funded grants. For example, in March 2003 the National Institutes of Health instituted a mandatory data-sharing policy. All NIH grants higher than $500,000 must include a plan by which final research data is made available to others for research purposes. IRs can be offered as a possible solution to this growing need.

Similarly, the authors of articles and monographs that require a lot of data manipulation receive requests for copies of their datasets so that others can either verify the original findings or use the data to support similar research. IRs are an easy way to accomplish this sharing. And, if the author makes the deposit before printing, the persistent URL to the datasets can be included in the publication itself.

Examples of datasets in IRs:

- Case Western Reserve University's The Common Folk Oral History Collection—https://tech-team-help.cwru.edu/handle/1763/16

- Erasmus University Rotterdam's Database of Happiness—https://dspace.ubib.eur.nl/handle/1765/288

Supplementary materials

A common frustration of authors is that the economics of publishing force them to exclude material from an article or monograph. Some publishers place restrictions on the number of images, particularly color images, which can be included, or charge the authors for each image.

At times, authors have been forced to eliminate entire sections or chapters of a work to keep down the total page count. IRs provide an opportunity to share all the material that for whatever reason did not appear in the final, published version.

Online and overlay journals

One way in which IRs can challenge the current scholarly publishing paradigm is through the creation of online and overlay journals. Numerous quality, peer-reviewed, online journals have been created over the past decade.

Often single individuals or departments at a university, with limited technical support and resources, undertake these enterprises, such as the online journals *Invisible Culture* and *Passages: Interdisciplinary Journal of Global Studies,* both of which are currently published through the University of Rochester's website.

IRs offer significant benefits to online journals such as these by providing a modern infrastructure and distribution mechanism, storage and backup systems, and technical support. By placing their online journal within the IR, the journal's editors can focus more on the journal's content and less on the technical infrastructure behind it.

Examples of online journals in IRs:

- *Dermatology Online Journal*—http://dermatology.cdlib.org

- *Engineering and Science* (in progress)—http://clsdemo.caltech.edu

An overlay journal is a collocation of links to articles that physically reside in distributed archives. The papers are deposited into various archives and then brought together virtually under the umbrella of the overlay journal. An IR can provide an overlay journal with repositories from which to select appropriate material, or it can host the overlay journal itself.

Examples of overlay journals:

- *Algebraic & Geometric Topology*—www.maths.warwick.ac.uk/agt

- *Annals of Mathematics*—www.math.princeton.edu/~annals

Books

Until the point at which the manuscript is printed, a book is in digital format. The growing corpus of electronic books is due in large part to the existence of these digital files, so naturally collections of electronic books are beginning to appear in IRs.

In the case of MIT, the collection consists of out-of-print titles from MIT Press. A university press, however, is not the only source of digital books. Usually, once its publisher has declared the title out-of-print, the copyright reverts back to the author. Thus, at a large academic campus, potentially hundreds of out-of-print book titles exist whose copyrights reside in the hands of faculty.

Ideally, the author would be able to obtain a digital copy of the final, formatted version of the book from the publisher. In the absence of this access, however, a less-polished version is still of value and use.

In January 2004, Cornell University launched Internet-First University Press, which will distribute the full-text of new books for free through the university's IR. Readers can download and view the texts online or exercise the option to purchase a print edition through a print-on-demand system.

With this open-access book publishing model, Cornell University hopes to "reduce the financial risk for the publisher by eliminating the need for a large inventory," while leveraging the benefits of wide, unfettered electronic distribution.

Examples of book collections in IRs:

- California Institute of Technology's Books by Caltech Authors—http://caltechbook.library.caltech.edu/

- Cornell University's Internet-First University Press—http://dspace.library.cornell.edu/handle/1813/62

- MIT Press Out-of-Print Books—https://dspace.mit.edu/handle/1721.1/1776

Learning objects

Although learning objects can be large, complex files, they are no less in need of proper care, storage, and distribution. The success of the Multimedia Educational Resource for Learning and Online Teaching (Merlot), which is a registry of online learning objects, demonstrates the tremendous interest in sharing and reusing digital objects. But for learning object registries, including Merlot, to succeed, permanent, persistent storage, such as IRs, must be available to house the actual learning objects.

Institutions with online course management systems, such as WebCT and Blackboard, are probably already aware of the need to store learning objects to be reused in future semesters. An IR may be a potential storage and distribution device for learning objects, if the system can provide for deposits of large, complex content objects.

If your organization has a strong interest in creating learning object collections, then you may need to look for an IR system that is compliant with the Sharable Content Object Reference Model (SCORM).

Examples of learning object repositories:

- Arcadia University Learning Object Repository—http://courseware.acadiau.ca/lor/index.jsp

- University of Arizona's DLearn—https://dlearn.arizona.edu/index.jsp

Multimedia collections

Scholarly communication is not confined to textual documents. The gray literature of some disciplines, such as art, music, and film studies includes multimedia formats as well.

Images, video, audio, and other multimedia content of enduring value are ideal content for IRs because they tend to be more ephemeral in nature than text documents and, therefore, in greater need of stewardship.

Multimedia collections, though, can be complex and require additional support, such as streaming server integration and thumbnail displays. Moreover, multimedia collections require special care to ensure all participants involved have given permission for the distribution of the performance.

Examples of multimedia collections in IRs:

- California Institute of Technology's Biological Imaging Center's Frog (Xenopus) Gastrulation Images—http://clsdemo.library.caltech.edu

- CERN's video collections—http://cdsweb.cern.ch/?c=Videos&as=0&ln=en

- Indiana University, Purdue University at Indianapolis' Herron School of Art Gallery—https://dspace.iupui.edu/handle/1805/41

- University of Rochester's 19th Century Hospital Images—https://dspace.lib.rochester.edu/handle/1802/34

- University of Toronto's OISE/TV (Educational Multimedia)—https://tspace.library.utoronto.ca/handle/1807/366

Electronic portfolios

A collection of electronic portfolios (e-portfolios) can be of particular value to students and faculty at any educational level. An e-portfolio is an excellent way for students to demonstrate their skills and acquired knowledge to potential employers and college admission boards.

Moreover, with the growing demand for demonstratable evidence of student learning, a portfolio of a student's work from convocation to commencement may be a powerful assessment measure. Faculty can use an e-portfolio collection to build online CVs, both for public consumption, as well as for tenure and promotion reviews.

Examples of e-portfolio collections:

- Portfolio @ Duke University—https://portfolio.oit.duke.edu/index.jsp

- University of Minnesota, Duluth's ePortfolio—http://eportfolio.d.umn.edu

Policies

Part of determining how an IR will be used extends beyond the question of the types of collections it will house and ventures into broader policy issues. Although you may not be able to predetermine all the relevant policies ahead of time, you should work through some of the larger policy issues early in the process, such as who is eligible to make deposits. Some policy decisions have technical implications you need to know before selecting an IR system.

Here are some issues your policies need to address:

- Who can make deposits into the IR? All members of the organization or just one class of members, such as academic faculty?

- Broadly speaking, what types of materials will the IR house? Just finished works or can drafts be included as well? Source material, including datasets, or just the documents that represent the analysis of those source materials?

- Can the IR be used for commercial purposes?

- Will the collections consist only of materials authored by members of your organization or can someone from your organization sponsor the material of others, such as the conference proceedings of a national organization or the newsletter of a local society in which a member of your board of trustees is involved?

- Will all material be welcome or just those approved by appointed people within the organization?

- If a person leaves the institution, what will happen to that person's content in the IR?

- Can material be withdrawn from the IR, and if so, how and under what circumstances?

- Will versioning be supported? If so, should one version replace another or will all versions be retained?

- Who owns the content once it is deposited into the IR?

- Is participation in the IR an option or a requirement, such as mandatory submission of an ETD?

- Will your access philosophy be more closely akin to "everything *open* to the world, unless compelling reasons otherwise," or "everything *closed* from the world, unless compelling reasons otherwise?"

- Who has the authority to decide when a new collection should be created?

- Will any preservation promises or guarantees be made for materials in the IR? If so, for all formats or just specific ones?

- Who is responsible for ensuring copyright compliance?

- Who is legally responsible for the IR?

You should not have to reinvent the wheel when developing your IR policies. The IR policies of another organization are a great starting point.

Your policy team could read the IR policies of an institution similar in size and type, and then decide whether they are applicable to your organization. Often just some small wording changes are all that is required.

Fortunately, many institutions post their IR policies on the Web, including:

- Archive of European Integration, http://aei.pitt.edu

- Florida State University, http://dscholarship.lib.fsu.edu/about.html

- MIT, http://libraries.mit.edu/dspace-mit/mit/policies/index.html

- Monash University, http://eprint.monash.edu.au/information.html

- National University of Ireland, http://eprints.may.ie/faqs.html

- Open University of the Netherlands, http://dspace.learningnetworks.org/index.jsp

- Simon Fraser University, http://ir.lib.sfu.ca/policies/community_and_collection.jsp

- University of Calgary, www.ucalgary.ca/library/dspace/policies.html

- University of California, http://repositories.cdlib.org/escholarship/policies.html

- University of Rochester, www.library.rochester.edu/index.cfm?PAGE=1285

- University of Melbourne, www.lib.unimelb.edu.au/eprints/collectionpolicy.htm

- University of Minho (Portugal), https://repositorium.sdum.uminho.pt/about.jsp

- University of Oregon, http://libweb.uoregon.edu/catdept/irg/AboutSB.html

- University of Toronto, https://tspace.library.utoronto.ca/policies/policies.jsp

- Vanderbilt University, http://www.library.vanderbilt.edu/dspace/
  Community_Collection_policy.html

Keeping your initial suite of policies in draft form is helpful. Since you cannot predict all potential uses of the IR, you may inadvertently turn away desirable content because of exclusionary policies. At the same time, you can quickly make modifications to justify excluding particularly messy or undesirable collections, such as a three-terabyte collection of geospatial data.

Services

In addition to the IR itself, many related, ancillary services can be offered with the IR, including digitization, metadata enhancement, batch importing or exporting, proxy submission services, and user support. Deciding beforehand which, if any, of these services you will offer could influence your selection of an IR system, as well as have a real impact on the projected cost of the project.

Although these supplementary services may seem to distract money, staffing, and focus from the central IR project, you will discover that many of these services can be deal breakers for recruiting content. Unless you are certain of ample material for the IR, particularly in the early trial period, you may have to offer some of these support services, else risk that many documents do not make their way into the IR.

You might decide to offer these services only to the earliest contributors to help build an initial critical mass in the repository. Or you may determine that some of these services will always be necessary to ensure the quality of the content and metadata in the repository.

In addition to whether to offer these services, you also have the decision of how to pay for them, which is addressed in Chapter 8.

Digitization

Potential IR materials are not always in digital formats, such as paper texts, photographs, slides, VHS, and 8mm films. Although the most recent papers in a work paper series may be in digital format, often the majority of a series' back issues are available only in paper.

If your institution does not already have one, you may need to set up a digitization lab that includes form-feed scanners, high-resolution flatbed scanners, and VHS to DVD conversion equipment, for example. Another option is to develop outsourcing contracts with local vendors. At the least, you should have a list of recommended local vendors.

Metadata enhancement

Without robust metadata, content could be lost amid a sea of digital documents. As any library cataloger can explain, however, the assignment of quality metadata is an art form, and one that requires much experience and training.

If the assignment of metadata is the sole responsibility of the person who deposits the document, then the quality of metadata will vary greatly throughout the IR. Take, for example, the importance of a controlled vocabulary. Without mandating the use of controlled subject terms and name authority

control, discovering items in an IR could become as hit or miss as finding documents using Google.

With this consistency in mind, you may decide to provide metadata services. On one end of the spectrum, this service could be consultative and only used when a collection is initially established. You would create a controlled vocabulary, determine which fields should be required, and decide on a metadata schema.

On the other end of the spectrum is a service in which a metadata professional reviews or generates the metadata for every deposit into the IR. MIT Libraries, for example, created a fee-based metadata service unit in part to provide metadata expertise for its DSpace institutional repository.

Batch importing or exporting

Batch importing can be a deal breaker when migrating large, pre-existing digital collections into an IR, such as a collection of several hundred working papers. Without some compelling reasons, the owner of the collection is not going to rekey all the collection's metadata to transfer it to the IR.

If the metadata can be automatically extracted, mapped to metadata fields used by the IR system, and then batch loaded, the advantages of the IR may be compelling enough for the content owner to move the collection.

Batch exporting can be valuable if data needs to be copied into other databases. In disciplines such as physics and economics, already well-established subject-based repositories exist. *Not* participating in arXiv and RePec is not an option for most physicists and economists, respectively, because these repositories have become a part of scholarly communication within the disciplines.

Faculty will probably not opt to deposit their documents into an IR *instead* of the established repository of their disciplines, nor would they be interested in taking the time to make deposits into both. With batch loading and exporting, however, you can provide alternatives that do not affect the work practices of the faculty members.

For example, a researcher could deposit a physics preprint into the IR, with the understanding that the document and metadata would be automatically copied and batch loaded into arXiv. An alternative scenario has the researcher depositing the preprint into arXiv, and an IR staff member harvesting it into the local IR.

In either case, the advantage to the researcher is that the preprint is easily discovered through arXiv, while being cared for and preserved by the IR. Moreover, inclusion of your IR's content in Thomson ISI's Web Citation Index and other third-party indexing services depends on the ability to batch export content.

Proxy submission

Although members of your organization may be excited about moving their digital collections into an IR, their enthusiasm may not be sufficient to sustain them through the process. An organization commonly establishes an IR because of a high level of interest among its members, only to find that few have the time to migrate their documents from departmental websites, personal Web pages, and hard drives into the repository.

Some departments can assign the task of document submission to an administrative or student assistant. Others may not have this luxury and look to the IR project for proxy submission assistance.

One word of caution—some IR systems include a click-through distribution license as part of the submission process, which places the burden of ensuring copyright compliance on the submitter. If you are going to offer proxy submission services, establish some means to ensure that your proxy submitters are not held accountable should your community members make poor or ill-informed choices when selecting documents to deposit into the IR.

User support

Unless you are working with a technologically savvy or relatively small organization, not offering at least some user support services will be difficult. User support can include services such as initial collection setup and customization of the user interface at a collection level. It may include technical support when submitters are struggling to uploading documents or would like advice on the best preservation formats to use.

Training of submitters and collection administrators also falls under user support services. As a general guideline, the level of user support required for an IR is probably quite similar to that of an online course management system, such as WebCT and Blackboard.