ELIZABETH BEDFORD

6

Stewardship

Academic libraries have always taken their duty to preserve traditional collections very seriously. But there is an increasing sense of urgency around our additional responsibility to steward the research output of our institutions—especially as digital scholarship transforms those outputs, and forces libraries to wrestle with new questions and new challenges. Stewardship is in many ways a Sisyphean task, requiring the ability to accept that "perfect" preservation is both a useful aspiration and an unattainable goal; we will never be "done" preserving an object, physical or digital. Faced with this potentially overwhelming reality, the question becomes how to identify what *is* attainable, what decisions can be made, and what actions can be taken.

This chapter will address three areas of uncertainty that provide framing for decision-making—or lack thereof—around taking digital materials under library stewardship. I have borrowed these areas from Janet Gertz's (director of preservation, Columbia University Libraries) excellent article "Should You? May You? Can You?," which was originally introduced to me through the work of the UW Libraries' Digital Workflows Task Force.[1] Broadly, Gertz differentiates among three kinds of decision-making in digitization projects: *should* concerns selection, *may* refers to legality, and *can* encompasses technical

From the Culture of Digital Scholarship in Academic Libraries, edited by Robin Chin Roemer and Verletta Kern (Chicago: American Library Association, 2019).

and financial resources. Although the article focuses these questions on the challenge of choosing physical materials to be digitized and made available online, the framework is much more widely applicable than this one mode of digital collections. In each of the following sections, I give background on the ways the three questions manifest when thinking about stewardship of digital scholarship in its many forms, providing illustration from University of Washington (UW) Libraries' experience in developing a new research data repository, DRUW (Data Repository at University of Washington). Our services at UW are a work in progress, and although some of our experiences with DRUW provide useful lessons learned, others reveal the questions that remain unanswered.

But first, let's establish some context for the challenges we face in stewarding the products of digital scholarship.

CHANGING ENVIRONMENTS FOR PRESERVATION AND ACCESS

Traditional journal articles and monographs offer significant advantages when considering whether libraries should, may, and can preserve the materials. For the most part, they are self-contained units of scholarship, produced by an established workflow of submission, review, and publication. They are widely recognized by academia as being fundamental to scholarly discourse, and their long-term value, either for the advances they contain or the record of scholarly thinking they represent, is usually uncontested. In the paper era, the preservation of these materials was widely understood to be libraries' responsibility. Crucially, these activities are explicitly supported by United States federal law; the doctrine of first sale (17 U.S.C. § 109) allows libraries to distribute single purchased copies of a work of intellectual property, and the libraries' and archives' exceptions to copyright law (17 U.S.C. § 108) allow libraries to make a limited number of copies of a work for the purposes of preservation and access. The shift in academic collections in the early 2000s from paper to digital did have significant repercussions for both preservation and access, because the move from purchasing to licensing and the US Copyright Office's clarification that digital display was protected under copyright meant that libraries' activities were no longer protected by the doctrine of first sale. However, traditional scholarly publishers, libraries, and other cultural heritage institutions have been able to take advantage of their long-standing relationships and workflows to create large-scale preservation partnerships. The journal preservation network LOCKSS (created and managed by Stanford University Libraries) and the nonprofit organization Portico (a service of Ithaka) are examples of this trend, as is the monograph-focused HathiTrust partnership.

The outputs of digital scholarship, in contrast, have none of these advantages. Unlike traditional articles and monographs, they usually do not have

From the Culture of Digital Scholarship in Academic Libraries, edited by Robin Chin Roemer and Verletta Kern (Chicago: American Library Association, 2019).

centralized publishing workflows that can be tapped as the basis for a preservation workflow. They are rarely "finished," but rather can be living works that are added to and edited over time. Their long-term value can also be ambiguous; some digital scholars may fight for their work to be recognized as perpetually valuable by the academic establishment, whereas others may feel that their outputs lose inherent value over time as techniques improve and new questions emerge. And they often incorporate potentially proprietary materials as objects of study in a manner and at a scale that introduces complex intellectual property considerations to their preservation. For libraries, the result is that the selection of digital research outputs for long-term retention is often unclear, legality is often questionable, and appropriate workflows are often opaque.

The Digital Challenge

If I want to figure out what's on a bunch of 3.5-inch floppy disks that were in a shoebox I pulled out of a closet, I need a computer with a disk drive and an operating system that can run the software designed to render the files. For most people, this would be a nonstarter—the current technical context has shifted so drastically from that of twenty years ago that this kind of setup would be extremely rare at home. In practice, all digital objects are created and used within a specific technological context, and as the technology landscape changes, our ability to preserve those objects for future use is therefore dependent on whether we can reconcile these different contexts. As advocates for future users, we face the further complication that we have no way of knowing precisely what those future landscapes will look like.

This is the puzzle that digital preservationists attempt to solve—how do you ensure this kind of reconciliation between contexts? To start, it's helpful to have a framework for what it means to have successfully preserved a digital object. Priscilla Caplan, former assistant director for Digital Library Services at the Florida Center for Library Automation, gives us an excellent breakdown of the elements of preservation, describing a fully preserved object as *available* (you have it), *identifiable* (you know what it is), *understandable* (you can interpret what's inside), *fixed* (it hasn't changed over time), *viable* (it's on media that hasn't degraded), *renderable* (you have software that can open it), and *authentic* (you know its provenance).[2] The preservation community has put a great deal of thought into the conceptual framework, individual pieces of information, and systematic workflows required for this kind of comprehensive stewardship.[3]

Unfortunately, the products of digital scholarship represent an extreme example of the challenges that can arise in the preservation process. Let's say a project takes the form of an Omeka website embedding a Twitter feed with Tweets that include a specific hashtag. Technically, this represents a complex

interaction between a sophisticated web platform (Omeka) and a database owned and hosted by a for-profit company (Twitter). We have techniques that can save and render the HTML and CSS code that underlies the website at a moment in time, but we can't just grab a copy of the entire Twitter database, which is what that code is designed to interact with. (The Library of Congress, in fact, made the decision in 2017 to vastly reduce its Twitter preservation program because the volume of data was overwhelming its resources.)[4] And even if we could, that copy would not reflect the changing nature of the database itself, with Tweets continually being added and deleted. Preservation works most easily with static snapshots, and the dynamic interactions that form the heart of many digital scholarship projects are extremely resistant to this kind of freezing-in-time. This is in fact one of the most significant and widely researched challenges currently facing the preservation community: how to wrestle with the logistical and technical questions of new kinds of output.[5]

Curation: Beyond Preservation

The concept of curation will be familiar to many in the archives and museums worlds, but it is also increasingly part of the conversation in the library world—particularly among data librarians.

Some libraries provide services to make sure researchers' data are well-documented and well-organized; rather than solely preserving a static file, these librarians work with the researcher to make sure that the file itself represents the most useful version of the data. This kind of service goes even further than Caplan's definition of understandability and is a growing area of interest for many information professionals. The UW iSchool associate dean for research Carole Palmer and colleagues describe this distinction as one between "data stewardship," which I would argue is much more aligned with the kind of preservation I described in the previous section, and "curation," which involves a more active interplay with the data itself: "Data stewardship is about management of a shared resource . . . but it is a function of 'managing data' that implies a less active, fixed maintenance of data over time. Curation, on the other hand, is concerned with availability and future use of data, including the enhancement, extension, and improvement of data products for reuse beyond a single scholarly community."[6]

The kind of service Palmer describes requires both a deep understanding of the disciplinary particulars of a data set and a deep investment of time for doing the work of curation. Think about a database of flu genome sequences on the one hand, and a set of images representing the output of materials science experiments on the other. The kind of description and organization required to make the two data sets understandable and useable is quite different, requiring adherence to different disciplinary metadata schema and

experience with different tools and software. In fact, the breadth of knowledge that would be required to service all of the potential kinds of data coming in to a domain-neutral repository would be extremely difficult to develop in a meaningfully comprehensive way.

Stewardship in the Real World

So, what happens when we simply don't have the time, energy, and resources to ensure all of Caplan's preservation elements, or to curate all of the data? One option is to throw our hands up in despair; the other is to take a deep breath and figure out what we can do with the resources in front of us. No heritage professional has ever said "I have too much time and too much money," and so the preservation community has in fact spent a lot of time thinking about what "good enough" preservation also looks like. The National Digital Stewardship Alliance Levels of Digital Preservation, for example, provide an excellent framework for those of us starting programs in less-than-ideal circumstances.[7] Shared infrastructure is another strategy; in the curation realm the Data Curation Network is an effort to address the issue of domain-specific curation needs by developing disciplinary understanding within one institution, while simultaneously creating a network to share curation duties with other institutions with different disciplinary expertise.[8] Stewardship exists along a continuum, and although we all aspire to the highest standards, it's important to remember that good work can be accomplished even if it doesn't perfectly address all of the existing needs.

From a practical perspective, the concept of *bit-level preservation* is incredibly important in resource-tight environments. The idea here is that if we keep the material as healthy as possible, preservationists and curators can have some expectation of success, if and when resources appear down the road. In Caplan's terms, we are attempting to make sure files are available, identifiable, fixed, and viable. Bit-level preservation involves a system that keeps track of the files and makes sure they aren't lost or degraded. This includes putting an automatic back-up process in place, and ideally it also incorporates hashing each file and carrying out periodic checksums to make sure that bit rot hasn't made the files unusable. This kind of preservation is significantly less labor-intensive, but make no mistake—stewardship in any form takes significant time and effort. In a world of finite resources, libraries must identify priorities for their research support services. The alternative—setting a goal of stewarding the research output of a university in its entirety—is a recipe for failure and waste.

This brings us to the first of our three questions for the stewardship of digital scholarship. Namely: should libraries take responsibility to address these vulnerabilities in the products of digital scholarship? And if so, where do we start?

1. SHOULD WE? LENSES FOR PRIORITIZATION

Historically, libraries have been in the business of building and stewarding collections. At academic libraries, those collections are meant to support, enable, and enrich the teaching and research carried out at individual institutions. And indeed, collections-related activities still form the bulk of the work and the majority of the expenditure at UW Libraries.

Yet there has always been an understanding that the line between collections materials and the scholarly output of a university can be blurred. Research is cyclical; the products of scholarship of one researcher form the basis for new work by another. Some of these products have long histories of explicitly being a part of the UW Libraries' collections policies. For example, our university archivist has long collected materials from faculty that include the notes, data, and objects that informed their published research. And in some ways, our traditional collection strategy of subscribing to departmentally relevant journals and monographs has already had the side effect of capturing many of the outputs of researchers on campus who choose to share their work in those formats. Nevertheless, the idea that UW Libraries has an explicit goal of collecting the research and scholarship produced by members of the University—in essence, that archival practices should be brought out of Special Collections and integrated into the work of public services at libraries—is quite a cultural shift.

An increasing community-wide interest in the idea of libraries as online publishing platform providers is supporting this shift. OCLC vice-president and chief strategist Lorcan Dempsey introduced the idea of the Inside Out library in 2010 and has subsequently developed the model into an extremely useful way of thinking about the purpose of a library.[9] Dempsey's viewpoint is that libraries should no longer solely focus on bringing external resources into a central accessible point—bringing the outside in—as was the strategy of traditional librarianship. Rather, he argues that libraries are in an excellent position to make sure that the resources that are currently under library stewardship are made as broadly accessible as possible—pushing the inside out.

The initial understanding of this strategy focused on unique collections and involved digitizing rare materials and making them available to a wider audience online. But over time, this understanding has morphed into a broader aim of providing stewardship services for the incredible unique resources produced by a university as a whole. The fact that materials produced outside of the traditional scholarly publishing workflow are particularly vulnerable make them prime candidates for library stewardship, which Clifford Lynch, executive director of the Coalition for Networked Information, articulates thusly: "the evidence suggests that the traditional commercial scholarly journals are pretty safe from a stewardship perspective, but the new components of scholarly communication . . . are typically very much at risk. These belong in institutional, disciplinary, and stewardship community repositories."[10]

Collections, Data, and the UW Libraries

But again, what does this mean from a practical perspective? As is the case with many large collections-driven research libraries, the UW Libraries' transition from "inward" stewardship and traditional collections to "outward" stewardship and online digital repositories has been shaped by a combination of inspirational values and challenging on-the-ground practicalities. For example, in 2014, the UW Libraries announced a three-year strategic plan that explicitly embraced both the idea of sharing collections with the world and the goal of making UW-produced scholarship accessible stating "we partner with faculty and students at the University to support not only discovery of and access to our collections, but we also aid in the management of scholarship and making accessible the exceptional work accomplished."[11] This strategic language is significant, but it is only the first step on the road to a functional and sustainable program.

Consequently, a number of our new projects have come about because of service-oriented librarians who see a need and then put in the legwork to determine if that hunch has broad enough implications to sway decision-makers in administration toward programmatic support. This was the case with our data repository project, which was given the green light based on the tireless advocacy of our former data services librarian, Stephanie Wright. Over the course of her time at UW, Wright developed close relationships with many scientists around campus. These interactions convinced her that a data-focused institutional repository, providing long-term archiving and access to data produced by UW faculty and researchers, would fill a pressing need.

Wright was responding to campus advocates of the worldwide open science and reproducibility movement, which has changed the conversation around sharing and archiving research data that underlie scholarly works. Broadly, the debate centers around two issues: scientists' ability to (1) repurpose others' data for their own, potentially radically different research, and (2) check others' claims by looking at the underlying data. This can be quite controversial for scientists who feel threatened by additional scrutiny of their research or who fear being "scooped" on potential publications by others taking advantage of their hard work. Interestingly, attitudes for and against are largely disciplinary based. The Inter-University Consortium of Political and Social Research, for example, has served the social science community as a centralized data archive since 1962. But many other disciplines are attempting to catch up, and individuals in those disciplines who are ahead of their peers in promoting those values need significant support.

A major driver for data sharing and archiving has been new funder requirements that explicitly address data management practices. A 2013 Office of Science and Technology Policy memo called on federal funders to create plans for promoting better data management practices among grantees, which has resulted in many funders now strongly encouraging open access to

the data underlying research results.[12] But this requires a platform for providing easy access to the data, and researchers whose disciplines do not have this infrastructure in place are left scrambling. This has led to a rise in interest for libraries-based institutional data repositories.

DRUW: Making the Case

Although Wright's intuition was important, launching a new initiative would require significant evidence of a broad need on campus. The UW Libraries has a strong culture of both assessment and user-focused services, and so one of the first projects Wright spearheaded after the creation of UW's Data Services program in 2012 was a Research Data Management Needs Assessment.[13] This work found "a significant need for research data storage solutions." Respondents designated the following as their top five priorities for libraries data-related services:

1. Ensuring that data is secure
2. Backing data up
3. Short-term storage (five years or less)
4. Long-term storage
5. Controlling and providing access to data

All of these services are well within the purview of an institutional data repository.

On the basis of this and other assessment projects, the UW Libraries decided to embark on creating a disciplinary-agnostic, campus-wide data repository service. Ultimate responsibility for the project fell to the Digital Repository Working Group (DRWG), recently renamed the Repositories Steering Committee (RSC), which is comprised of a team of librarians and staff members representing expertise in IT, scholarly publishing, metadata, preservation, and collections.

Setting the Boundaries

User assessments have been useful in gathering a broad baseline of information about the populations at UW who desire specific support for research data and were crucial in spurring and maintaining administrative support for the development of a data repository. However, their usefulness was more limited when it came to setting specific boundaries on the scope and priorities for DRUW. The Blue Ribbon Task Force on Sustainable Digital Preservation and Access, a seminal project investigating the economic factors that affect cultural heritage institutions' ability to steward digital materials for

the long-term, declares that one of the most important factors for ensuring sustainability is "a process for selecting digital materials for long-term retention"—in essence, creating a collections policy.[14] Without setting boundaries on a preservation effort's scope, it is extremely easy to become overwhelmed by the breadth of possible needs, and scarce funding and effort are in danger of being spent in ways that are ultimately less impactful.

As a scholarly publishing outreach librarian, I built off of the work of our former repository librarian, Mahria Lebow, and our current preservation librarian, Moriah Caruso, to lead RSC's latest efforts to create and confirm DRUW's policies. Developing an initial collections policy was extremely difficult, especially because the definition of data is so broad, and the long-term value of data is seen as wildly different among different disciplines. Complicating matters further, the UW Libraries does not have a central collections policy, but rather a template for individual liaison librarians to follow in creating their collections strategy documentation. As such, collections strategies at the UW Libraries are for the most part highly tailored and dependent on the expertise of individual liaison librarians. By contrast, for a disciplinary-neutral repository like DRUW to succeed, it must maintain a balance between the broad swath of user needs across the institution and the practical boundaries that ensure the repository's sustainability. It is a task that we have found extremely difficult, especially without the benefit of an internal collections precedent.

And so, RSC focused on the definitions we felt comfortable articulating. One boundary RSC was able to set has been around a very basic definition of "data:" DRUW's focus is on the objects of research rather than the products of research. One way of thinking about this is the distinction between scholarly claims (products) and the materials that allowed a scholar to come to and support that claim and their subsequent conclusions (object). In the heritage community, this split is often described as primary source (object) versus secondary source (product). In our collections policy we have therefore articulated the major qualifiers as the following:

1. DRUW is an archive for data that is an output of research conducted by University of Washington faculty, staff, or researchers. DRUW imposes no disciplinary restrictions.
2. If the research produces software and algorithms used to process data, or other protocols surrounding data collection and analysis, this may be contributed alongside any UW-produced or -modified data.

Our feeling was that we will have a better sense of whether we need to tighten our scope only after we begin to see what kind of use the repository is getting. Essentially, we are embracing iteration; to this end, we have set up a schedule for policy review that will force us to revisit the collections policy in two years, by which point we hope to have much more information about how the

repository is being used. In the interim, we have written a withdrawal policy, which leaves the UW Libraries the wiggle room to deaccession data sets if their inclusion negatively impacts the sustainability of the repository. Our question of "should we" steward the research data of the University prompts a broad response of "yes!," but we feel that we must gather more information before refining the practical details.

2. MAY WE? LEGAL AND POLICY IMPLICATIONS

Determining whether materials' content and value place them under the umbrella of libraries' collections policies is of course only the first step in making a determination about a stewardship workflow. We also need to be aware of the legal and policy context within which our programs operate, and whether those constraints affect the range of activities in which they can engage—particularly regarding intellectual property and privacy laws.[15] Digital scholarship materials often introduce specific legal challenges, having to do with an increased tendency for scholars to incorporate others' intellectual property into their own works, either linked or directly incorporated, and with the challenges that research data bring in terms of human subjects data and the tendency of many data scientists to recombine and reuse each other's data sets.

In the context of stewardship, distribution, or providing access, is extremely important when it comes to determining whether permission from the rights holder is needed or whether existing exceptions to copyright law cover libraries' activities. At UW Libraries, our preservation replacement program, for example, sends deteriorating physical books to vendors to scan and print new hard copies, which is covered in the exception to copyright protections detailed in section 108 of U.S. Code. In this process, the vendor also creates a digital file, which it returns to us alongside the physical book. We send this file to HathiTrust, which limits access for files that are still under copyright. Only if the copyright of these books lapses will Hathi add the digital files to its broad access program; we in the UW Libraries do not attempt to get explicit permission from the copyright holder to make the books available immediately. The end result of a workflow—whether it is solely for preservation or includes an aspect of providing access—is therefore a critical determinant in the kinds of legal and technical protections that must be put in place when designing a program.

For all digital products that libraries take into their collections with the aim of providing both preservation and access, there are a number of basic needs that must be met:

Permissions suitability. First, it is imperative that donors have the right to give us the material for the purposes of making it available openly. This means that they have all of the appropriate intellectual property

rights to enter into licensing agreements, and that the content itself is suitable for wide distribution—meaning the donors have explicit permission from human subjects and have excluded sensitive information and material restricted by intellectual property issues.

Library license. Second, there needs to be a mechanism for the donor to license the appropriate copyrights so that we are able to carry out the distribution and preservation of the materials.

End-use license. Third, whenever possible it is highly desirable for an end-user license to be attached to the material, to promote open scholarship.

Unfortunately, there are numerous complexities that libraries face in addressing these three requirements in the context of digital scholarship. Our experience at UW Libraries developing the DRUW Terms of Deposit illustrates this problem clearly.

DRUW: Getting the Rights

Data ownership and copyright is an ambiguous area in United States and worldwide intellectual property law. Facts, for instance, cannot be copyrighted, but the organization of facts (such as database structures) and the protocols for the collection and analysis of facts are potentially copyrightable and would likely be covered by intellectual property law. At UW, University policy states that under most normal research circumstances, copyrightable material is owned by the producer. However, when it comes to research data, the University asserts all ownership, a fact that would certainly surprise most UW researchers. In addition, our office of technology transfer asserts that commercial software is an exception to the copyrightable materials policy. Because DRUW is envisioned as a resource that will accept a wide variety of data—not only straightforward facts, but also copyrightable materials like software and images—UW's policies present a confusing mix of ownership issues that our repository's Terms of Deposit needed to address.

As I said earlier, it is crucial that materials in the repository are deposited by persons with the appropriate rights to do so, and that the library has the appropriate rights to distribute the materials if they are not already owned by the University. At UW, the *Grants Information Memorandum 37: Research Data* (GIM 37) provides the most direct insight into our University's current thinking on how data should be treated.[16] While drafting the DRUW Terms of Deposit, we initially attempted a blanket license from the University to the researchers that allowed them to deposit data into the repository. This was incredibly unwieldy, however, and we weren't sure it was absolutely necessary. GIM 37 explicitly differentiates among a number of roles when it comes to the treatment of research data: "Owner," "Stewardship," and "Access." UW

designates itself the owner of all research data, but the other two roles are delegated either to the Principal Investigator (PI) or to the PI's department. After consulting with UW's Attorney General's Office, we made the executive decision that by retaining the right to take custody of data, the memorandum explicitly gives UW, and by extension, the UW Libraries, the right to host data in the service of accessibility. By assigning a stewardship role to the PIs themselves, the memorandum gives researchers the right to deposit that data if they deem it appropriate. We therefore wrote the final Terms of Deposit to begin with a simple assertion that the PI's rights under GIM 37 allow PIs to deposit their research data.

This strategy does not, however, address materials owned by the researchers, and so the heart of the DRUW Terms of Deposit is a Grant of License from the Depositor to the University. Again, because we anticipate that future DRUW deposits will include copyrightable material that is owned by the researchers themselves, there needs to be a mechanism for the UW Libraries to get the rights necessary to distribute the materials. We therefore carefully crafted language to incorporate both researcher-owned materials and UW-owned materials:

> For all copyrightable materials in which the depositor has personal ownership, depositor grants UW a non-exclusive, irrevocable, worldwide license to exercise any and all rights under copyright related to the deposited materials and to their associated metadata, for the purposes of preserving them and making them freely and widely available in DRUW.[17]

Encouraging Open End-User Licenses

As a proponent and enabler of open scholarship, we at the UW Libraries are committed to encouraging depositors to license their materials using common open source licenses. Unfortunately, current UW policy did not make this licensing straightforward. We needed to first make sure that all stakeholders were on board with the idea of open licensing, and then figure out the mechanics of how to legally assign those licenses.

UW Executive Order (EO) 36 provides the basis for our university's policy on materials produced by staff or faculty that come under copyright or other intellectual property law.[18] The basic tenets of the policy are that under most circumstances copyright stays with the researcher, but that any patentable inventions must be assigned to the University or one of its appointees, and that licensing fees must be split between the inventor and the University in some proportion. The office that supervises these transactions is the University's Office of Intellectual Property and Technology Transfer, which is now called CoMotion.

The idea of openly sharing intellectual property was rather new to CoMotion, because they are primarily focused on helping researchers patent their inventions. In prior years, UW librarians had come up against some hostility to this idea, but the growing conversation among faculty in favor of openness has made CoMotion realize that having an option to share openly is extremely important. E036 as it is currently written is vastly too broad and does not lay out explicit instructions for materials intended to be distributed with an open source license. That said, a common sense reading of policy dictates that in a situation where monetization is not the goal, the only issue is whether any of the University's monetized materials would be in danger because of the license. In recent years, CoMotion has thankfully shifted its thinking to encourage more open source software licensing and has created both a guide for licensing open source software and a custom noncommercial use license for software. However, because CoMotion remains worried about UW-patented software, it explicitly forbids "Apache, GPLv3, other licenses that include a grant of patent rights which can impact the IP rights of the authors and other researchers at the University even if not involved in the project."[19]

Having established CoMotion's willingness to allow open source licensing, we then needed to determine the mechanism for applying that license. The question of who the licensor is for materials in a DRUW deposit, and which licenses will be offered, was nontrivial, because under US copyright law, only a copyright holder can license materials. Again, we know that materials deposited in DRUW may have complex ownership. Originally our strategy was for the PI/depositor to be the licensor, selecting from a predetermined list of possible licenses. But if that strategy were to be pursued, UW would somehow need to grant the depositor sub-licensing rights for materials that UW owned. Because GIM 37 does not mention sub-licensing as a responsibility with research data, and there are ambiguities about licensing in E036, we determined that it would be much more straightforward to have UW serve as the licensor, and have depositors grant the Libraries explicit sub-licensing rights for those materials subject to copyright. Consequently, we added a section to the Terms of Deposit that grants UW the right to sub-license any materials owned by the depositor based on the selection they make: "Depositor also grants UW a non-exclusive, irrevocable, worldwide license to allow others to exercise rights in accordance with a license selected by the depositor during the deposit process."[20]

Ultimately, we chose to offer depositors a broad range of open licenses to cover the broad range of potential content: Creative Commons licenses (https://creativecommons.org), an Open Data Commons license (https://opendatacommons.org/), the MIT license (https://opensource.org/licenses/MIT), and the BSD license (https://opensource.org/licenses/BSD-3-Clause). We set the Creative Commons Zero license (CC0) as our default setting because it is applicable for all content types.

Open Licenses

Creative works, software, and many kinds of data are protected under intellectual property (IP) law, which means that others who want to use those materials often need explicit permission from the IP holder to do so. A license is the mechanism that the IP holder uses to tell others what they can and cannot do with their work; open licenses allow the IP holder to preemptively grant permissions to others.

Open Source Software Licenses

The nonprofit organization Open Source Initiative (https://opensource.org/) keeps a list of licenses, which generally fall into two categories: those that are "viral," that is, require others to use the same license for derivative works, and those that are not. MIT and BSD are part of the latter category.

- SUITABLE FOR: software

Creative Commons "Zero" (CC0) License

The legal equivalent of designating something in the public domain—users may do whatever they want with the work, without crediting the author.

- SUITABLE FOR: software, data, creative works

Open Data Commons Licenses

A suite of licenses designed to accommodate the IP complexity that can arise with data sets.

- SUITABLE FOR: databases

Creative Commons Attribution Licenses

A suite of licenses designed to give IP holders a wide range of options for allowing and limiting reuse of their materials.

- SUITABLE FOR: creative works, some data

Settling on the mechanisms to answer the "may we?" question was a shockingly complicated and mind-bogglingly slow process. However, after two years and three rounds of negotiations, we now have confidence that our deposit agreement has the right balance of legal protection and encouragement for open sharing.

3. CAN WE? TOOLS, WORKFLOWS, AND STAFFING

The "can we?" question is where the rubber hits the road in terms of establishing a stewardship program. Ultimately, it doesn't matter if a library has decided that material is within its collections purview and that there are legal avenues for its collection and preservation, unless practical workflows for the ingest and stewardship of that material can be established and sustained. To do this, libraries must first define and articulate their responsibility toward these materials, and then explore the software and staffing necessary to put that responsibility into practice.

UW Libraries has explored a number of combinations of local or remote hosting and open or licensed software as part of its strategy to protect and provide access to its collections. Our existing platforms are:

- dSpace for our institutional repository, ResearchWorks—hosted locally, based on open source software
- CONTENTdm for our Digital Collections—hosted remotely as a service, based on proprietary software
- bepress for UW Tacoma's Digital Commons and the UW Law Library—hosted remotely as a service, based on proprietary software[21]

There is a robust body of literature around platform selection, particularly focusing on the trade-off between flexibility and effort required for maintenance of proprietary and open source software.[22] But a new and intriguing consideration is what role institutional values should play in selection. Bepress, for example, is an incredible platform, but its recent acquisition by Elsevier in 2017 has raised serious concerns about the trustworthiness of the company. The UW Libraries is still wrestling with the decision of whether to continue the relationship with bepress, in light of the challenges the alternatives would present.

The fundamental question of how much open source software costs is unfortunately tremendously opaque to most librarians. Libraries and archives have spent centuries honing the skills necessary to preserve and safely provide access to physical materials, and these skills are seen as firmly within the purview of LIS training programs. Digital archivists have been developing frameworks and strategies for their work since the 1980s, but only recently have the programming skills necessary to build and maintain digital preservation and access tools been conscientiously integrated into library curricula. Deteriorating relationships with for-profit vendors have made homegrown solutions extremely attractive. However, the IT skills gap has meant that it is easy for decision-makers to fundamentally miscalculate the time and effort such infrastructure represents. Open source software presents major challenges for development, and the UW Libraries' IT department was unprepared for what a difficult project it would be. Although our IT team was able to overcome

substantial hurdles to launch a pilot version of the repository, we are unsure about its long-term sustainability.

DRUW: Bit-Level Preservation with Hyrax

The DRUW project does not have assigned staffing for comprehensive preservation or curation, and so we made the decision that the service we could offer would simply be bit-level preservation. Still, this level of service requires the UW Libraries to create and maintain a platform to store and make accessible the deposited material, which in itself is no small task.

In the end, the Libraries decided to build the DRUW data repository with Hyrax software, an open source platform developed by the Samvera community that relies on Fedora as its underlying database.[23] We were excited about Hyrax partly because the Samvera community is engaged and enthusiastic, and partly because the kind of features that are being built into the program are direct reactions to issues that other repository software has surfaced. Online repositories essentially have three layers; they are a combination of (1) a web-based access platform, (2) an underlying database where the information is stored, and (3) a program that allows these two entities to talk to each other (often called the API layer). Fedora is an extremely robust database, and the initial vision for Hyrax was that the focus would be on making sure the middle layer could interact with multiple access layers that would be tailored to the content. Think about two repositories: one for student films, and one for traditional, PDF-based theses and dissertations. It would be useful for the film access platform to have streaming capabilities, but that would be redundant for the PDF-based content. Similarly, the kind of metadata that would be recorded for the two different content types would be considerably different. The idea is that Hyrax would allow separate access platforms for different repositories but only a single underlying database, which streamlines the technical architecture considerably.

Hyrax in Practice

Over the course of this project, as many libraries active in digital stewardship have discovered, we have learned that installing and configuring open source software can be remarkably difficult. Modern users of commercial software often don't understand that installation and configuration wizards are actually a miracle of development effort. In the same way, it is borderline magic that commercial software will send an automated notification that an update has been released, have you double-click on that notification, automatically make the update, and allow you to go on using the software without any

decrease in functionality. It represents the fact that many, many developers did a tremendous amount of testing behind the scenes to make sure that every potential feature still works properly after the update.

This kind of extensive testing doesn't happen with open source software, which means that every Hyrax upgrade comes with a high probability that some functionality elsewhere in the program will break. Furthermore, any customizations that our developers make on the local instance of Hyrax may or may not play well with the upgrade. So, every upgrade involves a significant amount of testing, tweaking, and potentially rewriting of previous customizations. As of writing this chapter, UW Libraries is able to devote the equivalent of approximately 1.5 full-time employees to its Hyrax development and maintenance (divided among at least five different real-life people). This number gives a clue as to why development of DRUW has been slow, and why we decided that our IT department cannot support customization. Instead, we will be focusing on maintaining a "vanilla" Hyrax instance and are trying to actively engage with the Samvera community so that our concerns can have a chance to be incorporated into the community development road map.

DRUW Development Skill Sets

Few people outside of IT departments have a deep understanding of the incredible breadth of skill sets required to develop and maintain an institutional repository hosted locally and based on open source software. The following five "roles" do not necessarily need to be played by separate people (although it certainly helps!); rather, they represent the kind of work that needs to be done, and for DRUW, the particular knowledge base that was required to create a Hyrax instance that would interact with UW's broader systems.

Information Architect

Decides how the system's schema should be configured to create an environment where deposited materials are organized and have appropriate descriptive information to be most useful in the future.

- SKILLS: Data modeling, metadata, workflow optimization

Developer

Customizes and configures the repository, including figuring out how to interact with other software and services.

- SKILLS: Ruby/Rails, Rake, Omniauth gem, Fedora, Solr, database management, CSS

(cont.)

Operations

Ensures that all of the software dependencies that comprise a functional repository environment are able to be installed and configured correctly.

- SKILLS: Vagrant or Ansible, Fedora, virtual machines, configuration management

Systems Administrator

Makes sure that the repository environment can be successfully deployed in the real world on existing hardware systems.

- SKILLS: Fedora, Solr, database management, Apache configuration, Shibboleth authentication, high availability server configuration, backup and security

User Experience

Tests that the repository functions correctly and identifies ways to make sure users' needs are being met and appropriate information is given to them.

- SKILLS: Quality assurance testing, user assessment, documentation development

Establishing Strategies for Risk Mitigation and Indemnification

Beyond the work needed to create the platform, RSC also needed to figure out how to make sure that no inappropriate materials get into the repository. Data collected through research on human subjects has the potential to contain sensitive information, which would be illegal to share openly. We also know that many UW researchers who reuse data often don't pay attention to the intellectual property status of existing data. If the data repository were to host such data, our technical infrastructure and quality control workflows would need to ensure compliance with both federal laws, such as HIPAA and FERPA, and university mandates.[24] DRUW was envisioned as a platform for completely open access, and so providing secure storage and controlled access to sensitive data would not only impose a significant burden on the system, it would also run contrary to the primary mission. And as I said earlier, current Libraries staffing isn't sufficient to review the data for compliance. So how do we address and attempt to mitigate the level of risk that our system unfortunately invites?

The Terms of Deposit ends with a section of warranties that do cover these issues; however, conversations with UW's Attorney General's (AG) office

revealed its opinion that relying on these warranties would not be enough to indemnify UW in case of a legal action. The risk is that in a lawsuit, depositors could argue that they hadn't read the Terms of Deposit and therefore didn't realize that they were not supposed to upload the materials. In order to combat this risk, the AG suggested that we implement workflows that would bolster the concept of *constructive assent,* whereby it would be difficult for a depositor to argue that they did not know or understand the limitations that were described in the Terms of Deposit Warranties section. As previously indicated, UW Libraries does not have the staffing capacity to create, implement, and sustain a new Hyrax feature that would pop up to remind users about sensitive data and intellectual property upon each deposit event, as the AG suggested. Some other workflow needs to be established, and we are actively exploring our options.

Ultimately, for DRUW, the answer to "can we"—and thus the answer to "will we"—is "we don't know." We have hard choices to make around whether the level of service we have been able to put together is in fact worth the effort it takes to maintain it. However, we have learned a tremendous amount throughout the policy and technical development of the project, and those lessons will be invaluable as we continue to develop our broader stewardship program.

CONCLUSION

I use the term "stewardship" throughout this chapter intentionally, partly because it has fewer connotations for heritage professionals than the relatively well-defined activities of preservation and access provision, and partly because it is something of an umbrella term that encompasses the full lifecycle of digital assets.[25] At UW Libraries, we are in fact wrestling with a lack of practical workflows and defined outcomes for our programs related to the preservation and access provision of nontraditional scholarly outputs. The term stewardship, for me, embraces this uncertainty while honoring a deeply felt belief that we have a responsibility to care for these materials and make them available to the world for the long term.

In this new and changing landscape, the trick is to make sure that we communicate our activities accurately and completely—that users have faith that we say what we mean and mean what we say. In particular, it's extremely important to remember that although librarians and other heritage professionals use terms like "archiving," "preservation," and "curation" as a shorthand among ourselves to designate a set of activities, these terms don't necessarily resonate with our patrons. In fact, they may have entirely different connotations depending on the circumstance. For heritage professionals, the term "archive" invokes a process of selection, description, arrangement, and preservation of materials, but for anyone using the Mac Mail app, the Archive

is the folder where users put emails they don't want to deal with but don't want to delete. And that's fine—most words have multiple definitions in the dictionary. We just need to make sure that we're clear to our users about which definitions we're using. Stewardship is a continuum, and part of our job is to help users understand and appreciate the steps we successfully take toward that long-term goal.

Takeaways

- A hallmark of digital scholarship is its evolving nature, and living resources present tremendous challenges in selection and workflows for preservation. Libraries must encourage disciplinary communities to develop norms around selection and retention and must also set and communicate explicit boundaries for themselves around stewardship responsibilities.
- Both the objects and products of digital scholarship are usually covered by intellectual property law, and therefore libraries must be extremely careful to make sure that they have the proper rights in place to steward the materials. Libraries should also encourage intellectual property holders to make their works available under an appropriate open license.
- Open source software development takes tremendous staff resources and requires extremely specialized skill sets. Libraries must be fully aware of these costs and have plans in place to fully support them in order to be successful in developing software-based services. If resources are unavailable, libraries have difficult decisions to make about how and whether to involve outside vendors in service provision.

NOTES

1. Janet Gertz, "Should You? May You? Can You?: Factors in Selecting Rare Books and Special Collections for Digitization," *Computers in Libraries* 33, no. 2 (March 2013), www.infotoday.com/cilmag/mar13/Gertz—Factors-in-Selecting-for-Digitization.shtml; Moriah Caruso et al., *Digital Workflows Task Force Final Report* (Seattle, WA: University of Washington Libraries, 2017).
2. Priscilla Caplan, "What Is Digital Preservation?," *Library Technology Reports,* no. 2 (February-March 2008) https://journals.ala.org/index.php/ltr/article/view/4224/.
3. The gold standard is the OAIS model, an incredibly useful and incredibly detailed examination of the activities that should go into a preservation program; see Consultative Committee for Space Data Systems Secretariat,

Reference model for an open archival information system (OAIS): Recommended practice (Washington, DC: CCSDS, 2012), http://public.ccsds.org/publications/archive/650x0m2.pdf.

4. Library of Congress, *Update on the Twitter Archive at the Library of Congress* (Washington, DC: Library of Congress, 2017), https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.
5. Those interested in digging into these questions have a huge body of literature to explore. Some good starting points are the journals *Code4Lib,* the *International Journal of Digital Curation,* the back issues of *D-Lib Magazine,* and *Journal of the International Association of Sound and Audiovisual Archives.* Relevant conferences include iPres, the International Digital Curation Conference, and the National Digital Stewardship Alliance's Digital Preservation Conference.
6. Carole Palmer et al., "Foundations of Data Curation: The Pedagogy and Practice of 'Purposeful Work' with Research Data." *Archive Journal* 3 (2013), www.archivejournal.net/essays/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/.
7. National Digital Stewardship Alliance, "Levels of Digital Preservation," https://ndsa.org//activities/levels-of-digital-preservation/.
8. Data Curation Network, https://sites.google.com/site/datacurationnetwork/home.
9. Lorcan Dempsey, "Outside In and Inside Out," *Lorcan Dempsey's Weblog,* January 11, 2010, http://orweblog.oclc.org/outside-in-and-inside-out/.
10. Clifford Lynch, "Updating the Agenda for Academic Libraries and Scholarly Communications," *College and Research Libraries* 78, no. 2 (2017), https://crl.acrl.org/index.php/crl/article/view/16577.
11. University of Washington Libraries Strategic Planning Committee, *Delivering Success: 2014–2017 Strategic Plan* (Seattle: University of Washington Libraries, 2014), www.lib.washington.edu/about/strategicplan/2014/directions/research-scholarship.
12. Executive Office of the President, Office of Science and Technology Policy, *Increasing Access to the Results of Federally Funded Scientific Research* (Washington, DC: Executive Office of the President, 2013), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
13. Stephanie Wright et al., *Fall 2012 Research Data Management Needs Assessment Results* (Seattle: University of Washington Libraries, 2013). See also chapter 8.
14. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation* (2008), http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf.
15. See chapter 3 on copyright.

16. University of Washington Office of Research, *Grants Information Memorandum 37: Research Data* (Seattle: University of Washington), https://www.washington.edu/research/policies/gim-37-research-data/.
17. University of Washington Libraries, *DRUW Terms of Deposit* (Seattle: University of Washington).
18. University of Washington Office of the President, *Executive Order 36: Patent, Invention, and Copyright Policy* (Seattle: University of Washington), www.washington.edu/admin/rules/policies/PO/E036.html.
19. University of Washington CoMotion, "Open Source: Releasing Software Under Open Source Licenses," https://comotion.uw.edu/what-we-do/intellectual-property-licensing/open-source/#section-1-0.
20. University of Washington Libraries, *DRUW Terms of Deposit*.
21. See chapter 10 for more about UW Tacoma's bepress work.
22. Two great starting points are Oya Y. Rieger, "Select for Success: Key Principles in Assessing Repository Models," *D-Lib Magazine* 13, no. 7–8 (2007), http://dx.doi.org/10.1045/july2007-rieger and Hillary Corbett and Jimmy Ghaphery, "Choosing a Repository Platform: Open Source vs. Hosted Solutions," in *Making Institutional Repositories Work,* ed. Burton B. Callicott, David Scherer, and Andrew Wesolek (West Lafayette, IN: Purdue University Press, 2015), https://crl.acrl.org/index.php/crl/article/view/16549.
23. More information on Hyrax may be found at https://github.com/samvera/hyrax/wiki. For more information on the Samvera community, see http://samvera.org/samvera-flexible-extensible/the-samvera-community/, and for more information about Fedora, see https://duraspace.org/fedora/.
24. The Health Insurance Portability and Accountability Act of 1996 and the Family Educational Rights and Privacy Act, respectively.
25. See the National Digital Stewardship Residency's stated mission of training professionals "in managing, preserving, and making accessible the digital record of human achievement." Library of Congress, "National Digital Stewardship Residency," www.digitalpreservation.gov/ndsr/.