2-1-2014

The Purdue University Research Repository: HUBzero customization for dataset publication and digital preservation

Carly Dearborn
Purdue University, cdearbor@purdue.edu

Amy Barton
Purdue University, hatfiea@purdue.edu

Neal Harmeyer
Purdue University, harmeyna@purdue.edu

The Purdue University Research Repository: HUBzero customization for dataset publication and digital preservation

Data sharing and open access are no longer simply buzz phrases in the scientific and publishing communities. Benefits to data sharing and reuse include increased collaboration, interdisciplinary innovation, and new solutions to pervasive social problems. The life and use of scientific data should extend beyond its original purpose. The new and sometimes competing demands placed on data have created what many call the "data deluge." Coupled with the sheer bulk of data created in modern research, the rapid advances in technology and tools, and the interdisciplinary research objectives can make open access a challenging objective.[1]

While the benefits of open access seem clear, the logistics surrounding sustainability are less so. A key factor in developing and sustaining open access to data is addressing issues surrounding preservation and data management. The emerging field of data curation combines the scalability of data management with the commitment to long-term preservation. An effective data management plan will factor in issues such as use of open standards for file formats, well-formed metadata, and information management literacy with the goal of viable future access.[2] The strategies of data management are currently being addressed more and more by university libraries and institutional repositories. These bodies are increasingly providing assistance with data creation, management, curation, and ultimately, preservation.[3] Purdue University Libraries specifically sought to operationalize this narrative by providing a platform on which Purdue researchers can receive data management support from subject librarians, fulfill the data management requirements of most funding agencies, take steps toward long-term data preservation, and provide immediate access to their research data.

Incentives and mandates for data sharing are changing the landscape of university libraries, especially at Purdue -- a leader in science, technology, and engineering research. More and more federal funding agencies require data management plans as part of their grant awarding process. In January of 2011, the National Science Foundation (NSF) began requiring a two page data management plan while other agencies have been requiring them for much longer. The National Institutes of Health required its grant applicants to take measures towards data management as early as 2003. The National Institution of Justice requires awardees to submit a data-archiving policy 90 days prior to the end of a funded project.[4]

The increased focus on data management and data sharing by these major funding agencies has necessitated a sea change in university library core functions. Librarians have the unique training to help researchers handle their data and prepare sustainable approaches to its management. It was this understanding that prompted the Purdue University Dean of Libraries, the Purdue University Vice

---

[1] Faniel, I. M., Zimmerman, A. (2011) Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *The International Journal of Digital Curation* 6(1): 59

[2] Lee, C., & Tibbo, H. (2007) Digital Curation and Trusted Repositories: Steps toward Success. *Journal of Digital Information*, 8(2). <http://journals.tdl.org/jodi/index.php/jodi/article/view/229/183>

[3] Cragin, M., Palmer, C., Carlson, J. & Witt, M. (2010) Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A, 368(*1926): 4023. doi:10.1098/rsta.2010.0165

[4] Witt, M. (2012) Co-designing, Co-developing, and Co-implementing an Institutional Data Repository Service. *Journal of Library Administration 52*(2): 3 doi: 10.1080/01930826.2012.655607

President of Information Technology, and Purdue University Vice President for Research to plan the development of a campus-wide data management platform using HUBzero software. This group of Purdue administrators created the Purdue University Research Repository Working Group in March of 2011. This group represented the major stakeholders from the University community and included experts from Purdue Libraries, Sponsored Program Services, and Information Technology at Purdue (ITaP). The Working Group began planning and developing the data management platform, now realized as the Purdue University Research Repository (PURR). Invested parties from the Libraries gradually formed the PURR Project Development Team (PURR Team). This team includes librarians, archivists, software engineers, and graduate students and is largely responsible for the continued development and maintenance of the repository.

This case study will discuss the progress of the Working Group and the PURR Team as they continue to develop PURR's platform and preservation infrastructure. This discussion will include the creation of guiding policies and procedures, plans to place those policies into action, and the unique metadata which describes and informs these actions. While PURR is operational, many of the components discussed within this case study are still in development and have not been fully implemented within PURR's environment.

CUSTOMIZATION OF HUB PLATFORM

PURR is a customized instance of HUBzero®, an open source software platform developed at Purdue University which supports scientific discovery, learning, and collaboration. HUBzero's concept originated from the US National Science Foundation's Network for Computational Nanotechnology (NCN). NCN's unique system, nanoHUB.org, allowed researchers, educators, and professionals to share resources and collaborate on issues surrounding nanotechnology. Built on the open source LAMP (Linux Apache, MySQL, and PHP) platform and the Joomla! content management system, nanoHUB allows for seamless integration of nanoHUB's many learning tools.[5] HUBzero is a generic version of nanoHUB's cyber infrastructure and is available for customization in other scientific disciplines. PURR is just one instance of a HUBzero customization.

The HUBzero platform seemed an obvious choice for the PURR Working Group. Its home-grown development allowed for easy code integration and provided a local technical support base. Also, its collaborative and publishing services matched the Working Group's strategic vision for PURR. PURR allows for collaboration within a *project*, or a dedicated working space. Projects provide space for data, wikis for project tracking and collaboration, and to-do lists. Once a producer creates a project, he or she can invite other collaborators to the project. PURR's platform allows producers to publish data for public access and discovery. A *dataset publication* is the term used to describe the published files and their associated metadata.

PURR DIGITAL PRESERVATION POLICY

[5] Klimeck, G., McLennan, M., Brophy, S.P., Adams, G.B., & Lundstrom, M. S. (2008). "nanoHUB.org: Advancing education and research in nanotechnology," *Computing in Science and Engineering, 10*(5): 17, 19, 22

While work on the technical infrastructure was underway, work also began an over-arching mandate for preservation of Purdue-affiliated datasets. A sub-group of the PURR Working Group, comprised of experts in digital initiatives, data curation, and digital preservation, co-authored the PURR Digital Preservation Policy.[6] The Preservation Policy provides a basic framework for the preservation direction and operations within PURR. At its core, the policy states that PURR, as part of Purdue Libraries, "is responsible for identifying, securing, and providing the means to preserve and ensure ongoing access to selected digital assets." A preservation priority structure was developed based upon dataset-to-publication relationship, ongoing teaching value, and long-term research value. The PURR Preservation Policy serves as an internal guide in conjunction with recognized digital preservation standards for preservation strategies and actions taken on PURR research datasets.

Part of the underlying foundation of the PURR service model is the promise of dependable, long-term storage of data publications. In order to confidently assure both the designated user community (Purdue's "faculty, staff researchers, and graduate students at Purdue University campuses and their immediate collaborators"),[7] and the larger academic community that PURR is a desirable, secure data repository, the PURR team charted a course to become a trustworthy repository via the Center for Research Libraries Trusted Repository Audit Checklist (TRAC) certification process. The TRAC assessment offered a roadmap to digital preservation infrastructure through its rigorous rubric. Within TRAC, considerations for metadata, software and hardware migration, digital object reliability (fixity), storage, disaster preparation, and much more are specified as integral to any well-formed digital preservation environment. The PURR team turned to ISO 16363 following its formal recognition—ISO 16363 is a direct descendant of TRAC and the international standard for Trustworthy Digital Repositories—as the guide for holistic preservation practices.[8] However, before any datasets could be ingested, code written, or preservation actions could take place, increasingly-granular policies had to be written, vetted, and added into the functional environment. Armed with ISO 16363 and internationally-recognized archival best practices, the PURR team began to build a digital preservation environment from the ground-up.

DEVELOPING POLICIES AND STRATEGIES

In early 2012, a digital archivist and metadata specialist joined the PURR Team to begin development of the policies that would inform the digital preservation strategies and actions of a fully-functional PURR. Taking into account the PURR Preservation Policy as well as ISO 16363, the digital archivist created two foundational preservation documents, the Preservation Strategic Plan and the Preservation Strategies.[9] The Preservation Strategic Plan outlines the overall objectives sought by the repository in order to

---

[6] The PURR Digital Preservation Policy was written by Paul Bracke, Associate Dean for Assessment and Technology, Jake Carlson, Data Services Specialist and Associate Professor of Library Science, and Sammie Morris, Head, Archives and Special Collections and Associate Professor of Library Science. The policy may be accessed on the Purdue University Research Repository website. https://purr.purdue.edu/content/article/51.

[7] PURR Designated Community definition: https://purr.purdue.edu/kb/AboutPURR/whocancreateanewproject

[8] Information regarding both TRAC and ISO 16363 can be found online at the Center for Research Libraries, *Metrics for Repository Assessment*. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying.

[9] As of July 1, 2013, these PURR documents are not available online.

provide long-term digital preservation. For example, one objective is designed to "ensure digital content is stored securely, redundantly, and remotely." However, preservation is useless without accessibility, so the objectives also reiterate that long-term accessibility is part of any fully-functional digital repository. As such, another objective is to "develop and implement strategies for planning for new formats and technologies."

Objectives are useful, but to accomplish anything, action is necessary. For each published dataset, the preservation activities documented in the Preservation Strategic plan must take place in order to ensure present-day and long-term use of the data. These activities are in accordance with archival best practice and relate to the provenance, intellectual content, metadata, archival information package (AIP), fixity information, and rights. The Preservation Strategic Plan also states that PURR "will be continually monitored and updated." Trustworthy repositories consistently make efforts to improve, and PURR is no different. In order to align itself with the most modern and comprehensive preservation techniques, PURR has and will continue to evolve its means and methods by seeking input from a variety of sources.

The second foundational preservation document, Preservation Strategies, details the strategies and actions each published dataset undergoes. For each dataset, PURR will undertake one of three preservation levels—bit-level preservation, full preservation, or no preservation. Nearly all datasets will undergo bit-level activities. Bit-level preservation will occurs if the object format is unrecognized or unsupported, or if full preservation is not possible at the time of ingest. In PURR the object file type is checked at ingest against the PRONOM file format registry using DROID (Digital Record Object Identification) to determine its level of support. (The DROID tool is discussed in greater detail later in this article.) Full preservation allows for the preservation of the original intellectual and structural content of the object. In full preservation, objects will receive bit-level treatment but also undergo transformation, normalization, or migration actions at ingest and whenever appropriate during the preservation life cycle. In extreme cases, datasets found to be corrupt or of potential danger to the repository will receive no preservation. Such objects will not remain within PURR.

More specific policies and preservation actions developed from the Preservation Strategic Plan describe categories of preservation support, preferred file formats, and the possible preservation actions for each. To illustrate, consider that a dataset using an open source file format which is supported by multiple software platforms is a better candidate for long-term preservation than a proprietary format with low adoption rates. The three levels of preservation within PURR will allow for flexibility at the point of ingest and during the life cycle of datasets. For example, a dataset ineligible for full preservation at its time of ingest may later become supported. New file formats, file support, and other new developments in the digital preservation field will continue. With the advent of new standards, PURR will adapt to strategically and programmatically implement new policies, strategies, and actions.

PURR's Preservation Strategies also lists the specific preservation actions which take place during the life cycle of a dataset in order to maintain its integrity and the renderability. These actions include: file format recognition, normalization, fixity checks, migration, and file backup.

File format recognition will take place once a dataset has been successfully published in PURR. At ingest each dataset is analyzed to determine its file format. PURR uses DROID, developed by The National Archives of the United Kingdom, to perform batch file analysis and identification. DROID, an open source Java tool, is able to provide detailed information such as file age, size, last modification date, or duplicate information. DROID then links that identification to a central registry, such as PRONOM, for technical information about each format.[10] PRONOM is also a product of the National Archives of the United Kingdom. Once the initial file format of the dataset is checked, this information will be documented for use in potential transformation, migration, and fixity checks during the entire life cycle of the dataset in PURR. Formats recognized by the DROID instance receive full preservation, while unrecognized formats undergo bit-level preservation until such time that the dataset may be further analyzed.

Normalization may take place to align the dataset with the current preservation standards. As each dataset enters PURR, items not structured within an established preservation file format may undergo a process to transform and normalize the object into an analogous long-term preservation format. For example, a text document will be normalized into a PDF. This process is undertaken to ensure the continued representation of the dataset throughout its life cycle. Normalization events are then recorded in preservation metadata (PREMIS) associated with the dataset throughout its life cycle.

Fixity checks are completed for all data within PURR on a regularly-scheduled basis to ensure no loss of data has occurred. Fixity checking is done by comparison of hashes generated at the time of ingest or after another preservation event. In the case a fixity check uncovers file degradation; the corrupted data will be removed and replaced with its uncorrupted counterpart at mirror sites. Hashes and fixity checks will be recorded in metadata associated with the dataset.

Migration to other preservation formats will take place over the life of many data objects. As such, PURR will continue to monitor content for potential file format obsolescence. If circumstances dictate data within PURR is at risk of obsolescence, the content will undergo migration to a new file format more conducive to its preservation. This will be done to bring PURR in line with rapidly evolving archival best practices and to ensure long-term preservation and access. This migration may include “upgrading” datasets to a newer version of the same format or transforming datasets into a completely new file structure. Migration events are also recorded in metadata associated with the object.

File backup is done for all content within PURR. All data within PURR is fully duplicated on a regular basis to prevent catastrophic loss of information. Purdue recently became part of the MetaArchive Cooperative and in time, information will be backed up and mirrored at other geographically-dispersed sites to provide a means of recovery in case of disaster. This is addressed later in this article.) File duplication also prevents data loss in cases of data corruption detected through regular fixity checks. In recovery events, actions taken will be recorded in metadata associated with the object.

[10] DROID. The National Archives of the United Kingdom http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm

Periodic checks of the datasets within PURR will determine what processes need to be enacted. Actions may be taken as many times as necessary to preserve the objects and datasets for the designated community. These preservation actions are the building blocks for the PURR preservation service model.

OAIS MODEL IN HUB ENVIRONMENT

Much like the customization of HUBzero into the PURR platform, PURR customized the Open Archival Information System (OAIS) Reference Model to fit its specific service model. An ISO international standard, OAIS provides a conceptual framework which defines the major functions involved in establishing an archival repository. OAIS was designed as a versatile and customizable model and does not suggest any particular implementation.

Image 1: PURR's OAIS Workflow[11]

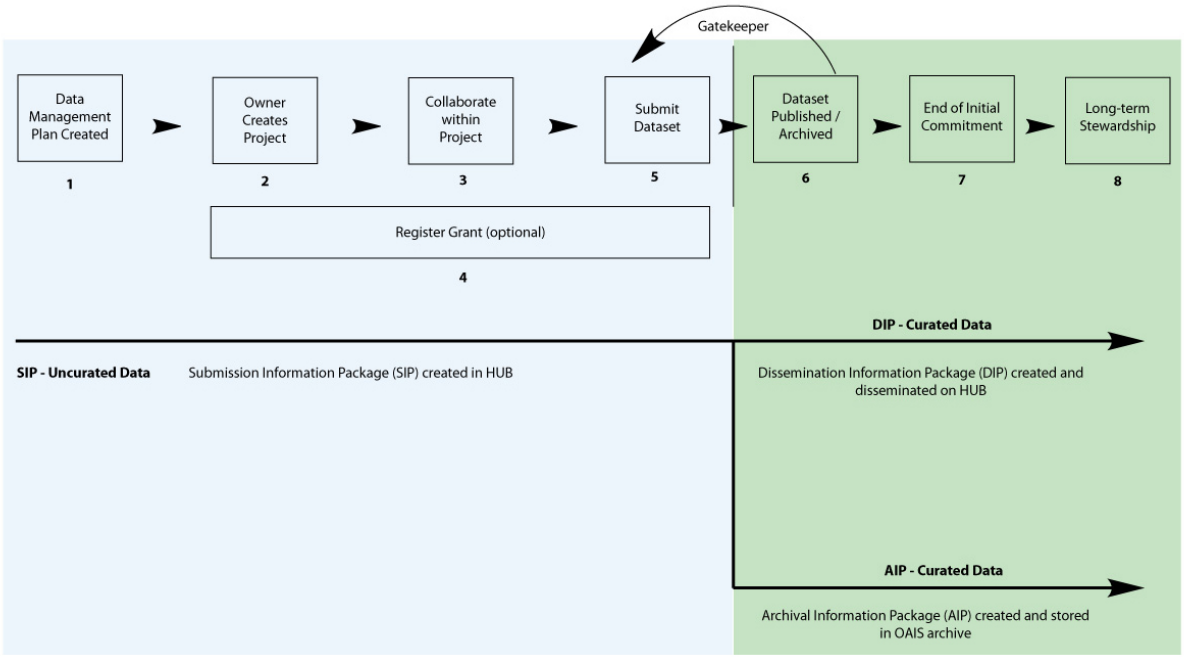

Image 1 shows PURR's designed workflow mapped to the OAIS Reference Model. The producer interaction with PURR begins with the creation of a project space. Prompted by a need for data management, preservation, or a wish to disseminate research, a producer can establish a Project Space on the PURR website (https://purr.purdue.edu), invite colleagues to collaborate, and eventually, choose to publish a dataset through PURR's online publishing platform. Once the producer is ready to publish a dataset, online forms prompt the producer to provide descriptive information for the dataset and to select a license for terms of use. This information in addition to the actual dataset comprises PURR's

[11] Image 1 designed by Brandon Beatty

Submission Information Package (SIP). The producer-supplied information is included in the descriptive metadata section which will be discussed later.

The final step of the publication process allows the producer to review and save the item or review and submit for publication. Selecting "submit" will mint a Digital Object Identifier (DOI), a unique and persistent identifier, which will facilitate data citation and access.[12] Selecting "submit" will also prompt the Gatekeeper function. This function represents the review process that PURR's repository specialist, or the subject librarian associated with the producer's discipline, conduct to verify that the dataset is an appropriate addition to PURR. The Gatekeeper reviews the SIP to verify that the producer provided adequate metadata, the content information is complete, and the entire submission meets with PURR's collection policy. If not, the Gatekeeper will send the submission back to the producer with comments for review or suggestions to place in another repository.

Once the Gatekeeper process approves the SIP, the process to create an Archival Information Package (AIP) begins. While the AIP creation tool has been written and tested, it is still in development and is not fully integrated with the live PURR site. PURR uses the Library of Congress BagIt specifications to package the dataset and its associated metadata – the descriptive information and the preservation description information. BagIt is a hierarchical file packaging format used primarily for storage and transfer of preservation-quality digital content. BagIt "bags" consist of a "payload," or the dataset encapsulated in the bag, and "tags," the metadata used to record bag transfer and storage.[13] The "bags" are read-only and cannot be altered once serialized.

CUSTOMIZED METADATA IN PURR

The HUBzero platform was not developed with metadata or preservation in mind. Therefore, a custom metadata implementation was necessary. after joining the PURR Team, the metadata specialist reviewed metadata standards to identify a "best fit" standard given the purpose of metadata in the PURR environment. The purpose is threefold: describe the dataset, identify dataset ownership and access conditions, and generate robust preservation metadata for long term preservation. At the conclusion of the standards review it was determined there was no single standard that met PURR's metadata requirements. However, by weaving together several standard schemes, all the requirements could be met. The metadata specialist developed PURR's metadata scheme using the following standards:

- Metadata Encoding and Transmission Standard (METS) [14]
- Dublin Core Metadata Initiative (DCMI) Metadata Terms (dcterms)[15]
- Metadata Object Description Schema (MODS)[16]

---

[12] In 2010 Purdue University Libraries became a founding member in DataCite, an international consortium which promotes the sharing of datasets by issuing DOIs. http://blogs.lib.purdue.edu/news/2010/02/16/purdue-libraries-a-founding-member-of-international-cooperative-to-advance-research/

[13] Boyko, A., Kunze, J., Littman, J., Madden, L., & Vargas, B. (2009) NDIIPP Content Transfer Project: The BagIt File Packaging Format, https://confluence.ucop.edu/display/Curation/BagIt

[14] METS. Metadata Encoding and Transmission Standard. Library of Congress. http://www.loc.gov/standards/mets/

[15] DCMI. Dublin Core Metadata Initiative. http://dublincore.org/documents/dcmi-terms/

- Preservation Metadata: Implementation Strategies (PREMIS)[17]

These standards, chosen due to their acceptance by the information management and academic communities, will continue to undergo support for the foreseeable future. More specifically, METS was chosen because it was designed to be extended by incorporating other defined metadata standards within the descriptive metadata and the administrative metadata container. METS acts as the wrapper into which the other standards are embedded. DCMI dcterms was selected for the descriptive metadata in anticipation of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) implementation within PURR. The MODS standard was selected to identify ownership and access conditions of the published dataset. Lastly, PREMIS was selected to support PURR's requirement for long term preservation of published datasets.

In order to achieve a high-level of dataset discoverability, the descriptive metadata must be complete and indexed. Datasets are described using the DCMI Metadata Terms (dcterms) standard. Through the project creation and dataset publication processes, the producer fills out online forms that capture descriptive metadata. The fields on the forms include Project Name, Project Alias, Title, Synopsis, Abstract, Authors, Tags, License, and Release Notes. The Project Name and Project Alias fields are populated by the producer at the time a project is created. During the publication process, the Synopsis field captures a succinct description of the dataset or the research that produced the dataset. The Abstract field provides more space for the producer to describe the dataset and/or the research project. The Authors field is repeatable and can capture a single author or the primary author and additional authors and/or contributors. The Tags field captures keywords that are indexed for dataset searching. The producer typically provides natural language terms in the Tags field.

Once submitted, the publication is queued until a subject librarian reviews the publication to ensure appropriateness, checks grammar and spelling, and adds controlled vocabulary subject terms in the Tags field. The natural language and controlled vocabulary terms in the Tags field, along with the other bibliographic fields such as Title, Author, Abstract, etc., enrich the description and are indexed for searching and discoverability. The License field is a dropdown list of Creative Commons[18] licenses reviewed and approved by the PURR Steering Committee. Finally, the Release Notes field is a place for the producer to include any notes with regards to file(s) descriptions, and any other pertinent information about the dataset or research methods.

Once the dataset is approved and published, the descriptive metadata values provided by the producer and subject librarian are stored in tables in PURR's database. The approval and subsequent publication of a dataset will eventually trigger an AIP Creation Tool to run. The Tool first process creates the AIP BagIt bag and then begins the process to dynamically generate serialize the metadata in a well-formed, validated Extensible Markup Language (XML) file to be included in the completed AIP. First, the METS

---

[16] MODS. Metadata Object Description Schema, Library of Congress. http://www.loc.gov/standards/mods/
[17] PREMIS. Library of Congress. Preservation Metadata Maintenance Activity. http://www.loc.gov/standards/premis/
[18] Creative Commons licenses. http://creativecommons.org/licenses/

metadata wrapper is written to the file. Next, the AIP Creation Tool maps the producer, librarian, project and system generated descriptive metadata values to dcterms elements and inserts the elements in the METS descriptive metadata container. The mapping is shown in Table 1.

Table 1: Field Mapping

PURR Field	dcterm Field	Value Generated By
Authors (primary)	<dcterms:creator>	Producer
Authors (secondary/contributor)	<dcterms:contributor>	Producer
	<dcterms:date>	System (ISO 8601 format)
Project Name	<dcterms:description>	Producer (project creation)
Project Alias	<dcterms:description>	Producer (project creation)
Publication State	<dcterms:description>	System (draft/review/published)
Publication Version	<dcterms:description>	System
Abstract	<dcterms:description>	Producer
Notes	<dcterms:description>	Producer
Synopsis	<dcterms:description>	Producer
	<dcterms:format>BagIt	System (AIP)
	<dcterms:identifier>doi	System (DataCite)
	<dcterms:publisher>	System
License	<dcterms:rights>	Producer
Tags	<dcterms:subject>	Producer/Subject Librarian
Title	<dcterms:title>	Producer
	<dcterms:type>dataset	System

The AIP Creation Tool will then generates the PREMIS preservation metadata for each file comprising the dataset. In the METS administrative metadata container a technical metadata section is generated for each file. Every file is a PREMIS object with a unique identifier. The dataset files are rendered read-only in the AIP so they may not be altered in preservation status. In the PREMIS metadata, "fixity-checked" and "read-only" behaviors are recorded as well as the content is a "primary" file or "secondary" file. Composition level, which indicates compression and encryption is determined and recorded. File size and format are identified and recorded in the size and format elements. As proscribed in the preservation documentation preservation level is determined based on file type recognition via DROID. If the file format is a supported format in PURRs preservation policy, the preservation level recorded in the metadata is "full." Otherwise, a file will have "bit-level" preservation recorded.

The PREMIS rights section follows the technical metadata section. PURR associates two licenses with a published dataset, the Terms of Deposit license and a Creative Commons license. The Terms of Deposit license grants permission for PURR "to use, duplicate and distribute the Work and to transfer the Work to any format or medium now known or later developed for archiving, preservation and access." [19] The

[19] PURR Terms of Deposit available at: https://purr.purdue.edu/legal/termsofdeposit

Creative Commons license is applicable to the dataset inclusive of all files therein. The AIP Creation Tool accesses the name of the license selected by the producer upon submission (e.g., CC0 - Creative Commons) and the location of the license terms (e.g., http://creativecommons.org/publicdomain/zero/1.0/) from the PURR database. In addition to the license information, the Tool assigns the first PREMIS Agent. According to the PREMIS Data Dictionary,[20] an agent is a person, organization, or software associated with rights management and preservation events such as file validation, repository ingest, migration, etc. Each agent has a unique identifier and is then recorded in the rights section. The license information is then written to the metadata XML file.

The next AIP Creation Tool process will generate the digital provenance metadata. The MODS standard is included in the provenance metadata to record the name of the primary contact of the dataset, a contact email, organizational affiliation, and, most importantly, the access conditions for the dataset. Upon publication, if the dataset is immediately accessible to the public, the value "publicly accessible" is recorded. However, PURR allows for a publisher to set an embargo with a specific release date. In the case of an embargo, the value will be, for example, "embargoed until 01/01/2014". It is important to capture this information for disaster recovery so that embargoed datasets remain inaccessible until the recorded release date.

The digital provenance section also includes all the agents and events that interact with the dataset and files. There are three agents recorded in the metadata. There is a Creative Commons agent that is the rights granting agent for the Creative Commons license. Another rights agent is the Terms of Deposit agent. The Terms of Deposit agent grants permission for the repository and PURR preservation staff to manipulate the dataset and files for purposes of long term preservation and access. The software agent, HUBzero, is also identified and recorded in this section. All agents have unique IDs so they can be linked to the events each performs on the dataset and files. Table 2 includes the events recorded in the digital provenance section

[20] PREMIS. Library of Congress. Preservation Metadata Maintenance Activity. http://www.loc.gov/standards/premis/

Table 2 PREMIS Events

Event Name	Event Description	Event Preservation Explanation
capture	Initial capture of the publication data from the user -- the first event in the event stream.	Preserving capture would help with HUBzero debugging, as it tells us when the SIP capture/creation process started.
in-revision	Generated when a SIP must be revised before it can be approved (it would occur between capture and validation).	Preserving in-revision would help with HUBzero debugging, as in-revision is a major change in the SIP status. Note that in-revision only occurs if the SIP is sent back to the author(s) for revision before it can be approved for AIP status.
validation	Validation of the SIP to ensure it is ready to become an AIP.	Validation is one of the major steps in a SIP's journey to AIP status, so it should be preserved.
ingestion	Creation of the AIP from the approved SIP.	Ingestion is the creation of the AIP, so it should be preserved.
fixity check	Periodic event where the fixity of the files in the AIP is re-validated.	Preserving fixity check would help with HUBzero debugging, as it tells when there is a problem with the preservation process.
replication	Copying the AIP bit-for-bit to another location for preservation purposes (as in LOCKSS).	As replication creates another copy of the AIP, it should be preserved for debugging purposes.
migration	Transforming the AIP and its contents into a more-contemporary format.	As migration creates a newer, automatically-generated version of the AIP, it should be preserved for debugging purposes.

21

The final process for the AIP metadata generation is the file section. The file section simply records the file hierarchy (primary, secondary, etc.), the files' unique IDs, and the files' storage locations. It also records the names of the files included in the AIP. After the AIP Creation Tool finishes generating the metadata, the well-formed, validated metadata is included in the AIP along with all the files and other preservation data and preservation files. The AIP will then be considered completed and ready for preservation.

In special occasions, an AIP can serve as a SIP. For example, in the case of transfer from one preservation system to another, a current AIP would serve as the submission package to this new platform. A Dissemination Information Package (DIP) is created from the same source material as the AIP; however, it is not a copy or a derivative of an AIP. A DIP is generated on demand once a member of the designated community visits the web interface and downloads the dataset. The designated community also has the ability to download the DIP's associated metadata files. This is not a typical process but one that works best for PURR's publication model, especially as the AIP workflow is still in development. Generating the DIP from the submission materials allows PURR to provide immediate access to published datasets.

CONCLUSION

While development of PURR's preservation infrastructure is ongoing, the team is making progress toward the goal of becoming a trusted digital repository. PURR will utilize a distributed digital

[21] Data Dictionary for Preservation Metadata, http://www.oclc.org/content/dam/research/activities/pmwg/premis-final.pdf

preservation model as a strategy for AIP back-ups. In early 2013, Purdue University Libraries became a member of the MetaArchive Cooperative. Developed in partnership between six southeastern U.S. university libraries with backing from NDIIPP, MetaArchive utilizes LOCKSS software to create a digital preservation network which approaches digital preservation through replication and geographic distribution. While still in the early phases of integration, MetaArchive promises to provide PURR with robust archival backup, in addition to Purdue's local and satellite storage infrastructure. Once Purdue and PURR are fully integrated with the cooperative, PURR will be able to satisfy additional ISO 16363 items.

ISO 16363 continues to serve as a rubric, barometer and set of goals for PURR as development continues. To become a trustworthy repository, the PURR project team has consistently worked to build a robust, secure, and long-term home for collaborative research. In order to fulfill its mandate, the project team constructed policies, strategies, and activities designed to guide a systematic digital preservation environment. PURR expects to undertake the full ISO 16363 audit process at a future date in expectation of being certified as a Trustworthy Digital Repository. Through its efforts in digital preservation, the Purdue University Research Repository expects to better serve Purdue researchers, their collaborators, and move scholarly research efforts forward world-wide.