Peer, L and Dull, J. 2020. YARD: A Tool for Curating Research Outputs. *Data Science Journal,* 19: 28, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2020-028

PRACTICE PAPER

YARD: A Tool for Curating Research Outputs

Limor Peer[1] and Joshua Dull[2]

[1] Institution for Social and Policy Studies, Yale University, New Haven, Connecticut, US

[2] Libraries, Collections, and Academic Services, The New School, New York, US

Corresponding author: Limor Peer (limor.peer@yale.edu)

Repositories increasingly accept research outputs and associated artifacts that underlie reported findings, leading to potential changes in the demand for data curation and repository services. This paper describes a curation tool that responds to this challenge by economizing and optimizing curation efforts. The curation tool is implemented at Yale University's Institution for Social and Policy Studies (ISPS) as YARD. By standardizing the curation workflow, YARD helps create high quality data packages that are findable, accessible, interoperable, and reusable (FAIR) and promotes research transparency by connecting the activities of researchers, curators, and publishers through a single pipeline.

Keywords: Data curation; reproducibility; data quality; workflow tool

Introduction

YARD (Yale Application for Research Data) is an adaptable curation workflow tool that enhances research outputs and associated digital artifacts designated for archival and reuse.

Quality and curation in repositories

The scientific principle of self-correction asks researchers to be transparent about their design, data, methods, and analysis. Transparency makes it possible for independent researchers to "reproduce reported results; test alternative specifications on the data; identify misreported or fraudulent results; reuse or adapt materials (e.g., survey instruments) for replication or extension of prior research; and better understand the interventions, measures, and context" (Miguel et al., 2014, p. 31). To the greatest extent possible, data and materials, including code, should be made publicly available in order to increase accountability for researcher error (Chambers, 2019; NASEM, 2018) and allow others to reproduce and confirm results (NASEM, 2019).

Universities, private and public funders, scholarly societies, journals, and other stakeholders in the scientific enterprise have been looking to data repositories to make research data, code, and other materials underlying reported findings more widely discoverable and accessible. Data repositories, however, do not apply uniform or standardized curation practices, with many offering self-deposit or opting for a minimal curation model in an attempt to appeal to busy researchers. So as the digital artifacts associated with research outputs proliferate via various repositories, they are not necessarily usable or interpretable. Stodden et al (2018) recently found "serious shortcomings in usability and persistence" of digital artifacts supporting publications (see also a discussion of attempts to evaluate empirical claims in published studies, Leek & Jager, 2017). We view the ability of future users to independently understand and reuse research outputs as a key aspect of quality (see also Altman, 2012; Ashley, 2013; King, 1995; The Royal Society, 2012).[1]

The cost of repositories' failure to ensure the usability and interpretability, or quality, of research outputs can be great. As data science emerges as the next frontier (Blei & Smyth, 2017; Burton et al., 2018), the ability to reliably use data and other digital artifacts associated with research outputs for the purpose of validating the integrity of scientific claims must be a precondition. From a practical standpoint, the community should expect that investment in research infrastructure is extended to the production of digital artifacts that can be

[1] Other aspects of quality, such as the accuracy and validity of the data or the soundness of analytic choices captured in code, are outside our definition of quality. We defer these evaluations to the scientific community.

used meaningfully. Moreover, given that, "a large number of scientific studies... suffer from the underlying computational and statistical issues" (Leek & Jager, 2017), usable and interpretable research outputs are imperative.

Curation of digital research data is traditionally defined as activities that reduce threats to their long-term research value and mitigate the risk of digital obsolescence (DCC, n.d.). We refer to gold-standard curation as the measures taken to ensure that research outputs are independently understandable for informed reuse (Peer et al, 2014). Some curatorial activities – such as the periodic review of the digital integrity of a file and remedial actions to protect data from digital erosion or hardware failure – need to be ongoing. Other activities – such as code review – may be more pertinent at certain points of the data lifecycle (see Johnston et al., 2014, for a comprehensive list). We believe all curation activities are vital.

Our concern here is with the optimization of curatorial activities that enhance the quality of research outputs and associated digital materials. Ensuring the quality of these digital materials designated for long-term reuse requires effort and inevitably some cost. It has been noted that, "managing research data for quality, in one form or another, has in fact been the core responsibility of data curation since its inception as a distinct sub-discipline within the library and information sciences" (Sposito, 2017, p. 3). At present, however, there is no consensus in the scientific community about who is responsible for this effort: Researchers, data centers, university libraries, or data repositories? Moreover, an analysis of curation practices in general purpose and domain repositories found that review and curation of these materials are sometimes minimal or limited in scope (Peer et al., 2014), with self-archiving models often guaranteeing no more than bit-level preservation in order to control costs.

Absent a consensus on responsibilities and practices, it is not surprising that there are very few tools currently available for curators and others to manage, standardize, and share responsibility for curation activities.[2] In contrast with custom tools built and used by individual repositories to accommodate their own specific curation needs and preferences, a universal tool affords other entities a way to engage with curation activities earlier in the data lifecycle. For example, laboratories can use the tool during active research, for example during data collection, to ensure that subsequent transformations to raw data are documented in sufficient detail. This can enable other researchers to trace the final analysis datasets supporting published findings back to those original raw data.

We describe YARD, a tool that responds to this challenge. YARD, the Yale Application for Research Data, is a workflow tool that facilitates gold-standard curation tasks. Our goal in developing YARD is to create highly curated research packages that can then be deposited into any data repository. In addition to the standard curation activities, the tool facilitates tasks for reviewing and enhancing research outputs, including verifying that data, code, and other relevant digital artifacts computationally reproduce the results those materials reportedly support. The tool also creates rich metadata about the artifacts, helping generate findable, accessible, interoperable, and reusable (FAIR)[3] digital objects. By tracking curation tasks, YARD supports a transparent and documented workflow that can help researchers, curators, and publishers share responsibility for curation activities through a single pipeline. And finally, by building flexibility into the system, YARD is designed to be adapt to changing requirements and standards.

General Description

The curation tool is a web-based application designed to process digital artifacts associated with research outputs, including metadata management, and to deliver highly curated research packages into designated repositories (see **Figure 1**). YARD is the implementation of the tool at Yale University's Institution for Social and Policy Studies (ISPS) which began in 2018.

Specifically, the curation tool offers two main benefits:

1. Managing complex workflows. The workflow design helps guide depositors and curators through tasks for reviewing and enhancing research outputs. The tool tracks these curation tasks and generates rich metadata. The tool can be used to manage any updates to metadata and data, which can then be pushed out to public repositories. An advantage of the tool is that the high quality data packages it produces can be linked to different endpoints for dissemination. For archives and repositories that already do a fair amount of curation, the tool facilitates a systematic workflow with tracking and integration capabilities. For self-archiving systems that offer little or no curation, the tool can be an option for depositors, as a means of enforcing minimal documentation

[2] We note the Dataverse Data Curation Tool, currently in development, which is meant to be used with Dataverse. See: https://github.com/scholarsportal/Dataverse-Data-Curation-Tool (accessed 2020 January 2).

[3] See: https://www.go-fair.org/fair-principles/ (accessed 2019 November 27).

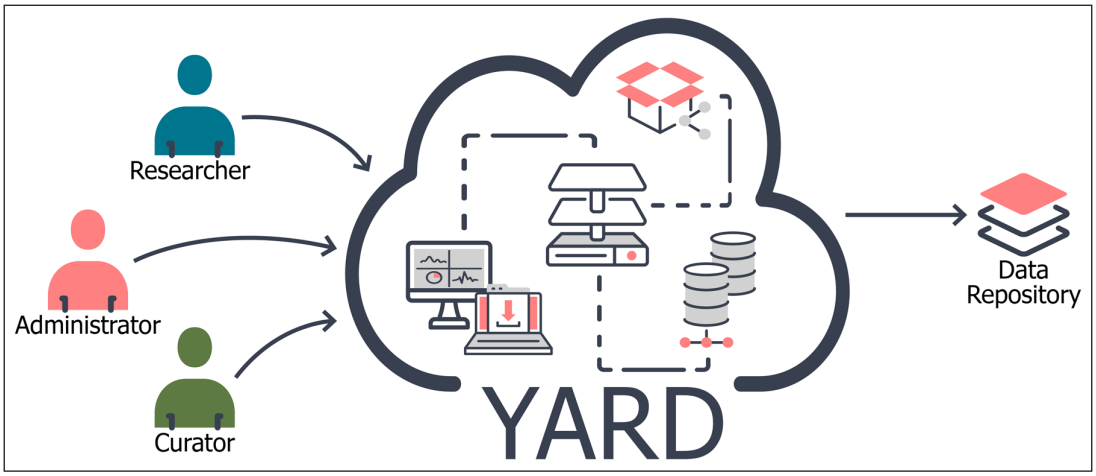

Figure 1: YARD is a workflow tool for reviewing and enhancing research outputs and delivering them into a repository.

standards, for example. In addition, organizations with distributed expertise can use the tool to collaborate, coordinate, and standardize curation activities. For example, university library staff may be responsible for metadata generation, a statistical support unit responsible for verifying computational reproducibility of statistical analyses, and a repository responsible for assigning persistent links. An advantage of a distributed workflow model is the potential to increase the feasibility of scaling curation services without shouldering the entire cost of labor and technology.

2. Enforcing data curation standards. Specifically, the curation tool supports FAIR principles for findable, accessible, interoperable, and reusable data and other digital research artifacts. It can accommodate extending curation workflows to include additional quality checks (e.g., verification of computational reproducibility). The tool supports archival preservation policies by enforcing standards (e.g., the OAIS Reference Model requirement to clearly define roles, see CRL, 2007) and providing documentary evidence of such, which ISO 16363 (ISO, 2012) and CoreTrustSeal require (Dillo, 2018). As standards evolve, the tool can be configured to adapt.

Technical Features

Key features are based on services critical to rigorous data curation: Templates for multi-file metadata creation and editing, item-level metadata creation and editing, metadata error reporting, customizable metadata exports, controlled vocabularies for selected fields, controlled vocabulary editing capabilities, record versioning, user access options, administrator and tracking controls, and a variety of content management features. The curation tool is API-enabled and modular.

For a minimal installation, the tool requires two open source software pieces, a web server, a database, and file storage.

The two open source software pieces, available on Github under a GNU Affero General Public License v3.0. (Iverson & Smith, 2018), include the Curation Service and Curation Web application.

a) The Curation Service manages the curation workflow and logs all application events. The workflow is based on established curation steps triggered by certain user actions, and automated when possible.
b) The Curation Web Application provides a web interface for the Curation Service. All users – researchers depositing data and code, curators processing the outputs, and administrators – can access the curation tool through the web application.

The curation tool also requires,

c) A web server to host the curation tool. For the YARD implementation, the Curation Service and Web Application are installed on a Windows 2012 web server, which also hosts other software components.[4]

[4] The components are written in C#. The current version of the curation tool runs on a Windows server and the Web framework is .aspnet.

d) A curation database for storing the application and curation metadata.
e) File storage for data, codebooks, code, and all other files required for curation. The application requires storage locations for phases, a) an 'original' directory for the original files and metadata comprising the research package, b) an 'active' directory for copies of the files and metadata during active curation, and c) a 'processed' directory for copies of the files processed and approved, as well as metadata.

The curation tool affords easy integration with other software or workflows. These optional components are not integral to the functioning of the curation tool but are congruent with its purpose; they can be replaced, enhanced, or left out per organization policy. **Figure 2** illustrates the curation tool components, their function, and relationship.

At Yale, the YARD implementation of the curation tool integrates with components that provide additional desired functionality. It is configured to the requirements of ISPS and includes some proprietary software components. The YARD implementation includes,

a) Colectica Repository, a proprietary software developed by Colectica to create variable-level metadata extracted from SPSS, Stata, CSV, and Excel files (Colectica, 2016). The metadata scheme is based on the Data Documentation Initiative (DDI)[5] (the tool allows adding new fields from other established metadata schemes). This software requires its own database to store the metadata.
b) ClamAV as an antivirus check for all uploaded files (Tiesi, 2014).
c) StatTransfer for creating plain-text copies of data files (Circle Systems, Inc., 2017).
d) Yale's Persistent Linking Service to create persistent URLs (Yale University, 2016).

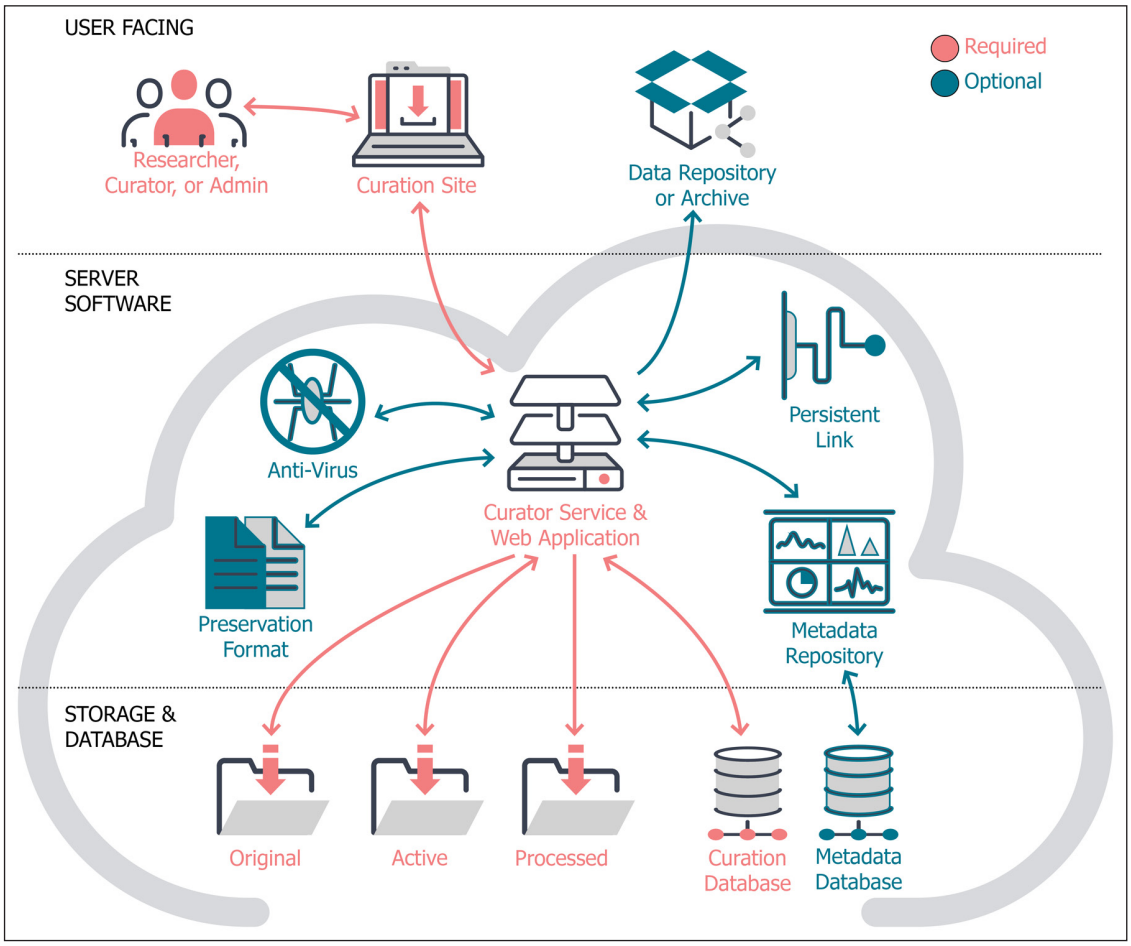

Figure 2: Curation tool components, required and optional.

[5] See: https://ddialliance.org/ (accessed 2019 December 30).

Table 1: Curation Tool Components.

Component	Function	License	YARD implementation	Alternate Component options
Required components				
Curation Web Application	Web interface for the Curation Service	AGPL 3.0	Curation Web Application	
Curation Service	Data deposit and curation	AGPL 3.0	Curation Service	
Curation Database	Storage for curation tool data	Proprietary	Microsoft SQL	Postgres, MySQL
File storage	Storage for files	Yale Service	Network attached local service (storage@Yale)	Any file storage (requires read/write access)
Optional components				
Metadata Repository	Captures, generates, and versions DDI Lifecycle metadata	Proprietary	Colectica Repository	Repository software that generates metadata
Metadata database	Stores variable-level metadata	Proprietary	Microsoft SQL	Postgres, MySQL
Anti Virus	Virus scan for deposited files	GPL	ClamAV	Any Antivirus software
File Conversion	Creates csv copies of data files	Proprietary	StatTransfer	Any statistical or custom software
Persistent Link	Persistent Link	Yale Service	Yale Handle service	Any persistent linking service

Table 1 lists the required and optional components, specifies the components used for the YARD implementation, and suggests alternative options for components where available.

User Roles

All users are required to create an account in the curation tool.[6] Users are assigned one of the following roles: Depositor, Curator, or Organization Admin. Users of the curation tool will experience a different workflow and have access to different features depending on their role in the system. Below, we discuss the curation workflow through the lens of the three main roles.

Depositor

By default, all users are Depositors. A Depositor can submit data, code, and other research outputs that comprise a Catalog Record. A Depositor can be a researcher affiliated with an organization hosting the curation tool or one of the organization's staff members. A Depositor creating a new record will add study-level metadata (e.g., author, title, sample size, field dates, etc.), upload all related files, add file-level metadata, and finally submit for curation. **Figure 3** is an example of uploaded files associated with a sample study.

Each file is checked by Clam AntiVirus, assigned a universally unique identifier (UUID), and deposited into the 'original' directory, where copies of all the raw data are kept. A Depositor can update the record if there are changes or new additions to a study and re-submit for curation. Each version submitted is preserved in the 'processed' directory so no data are lost.

Organization Admin and Curator

Once a Depositor submits a study for curation, copies are made and stored in the 'active' directory and a notification is sent to the Organization Admin. As noted, the Organization Admin has permissions to edit the organization's settings, including setting the domain name, assigning storage locations, specifying a repository destination, adding a deposit agreement, managing roles and permissions, and other technical settings. The Admin assigns a Curator to each Catalog Record and approves publication once curation is complete.

[6] The YARD implementation uses a password log in method; future development can include other methods of authentication.

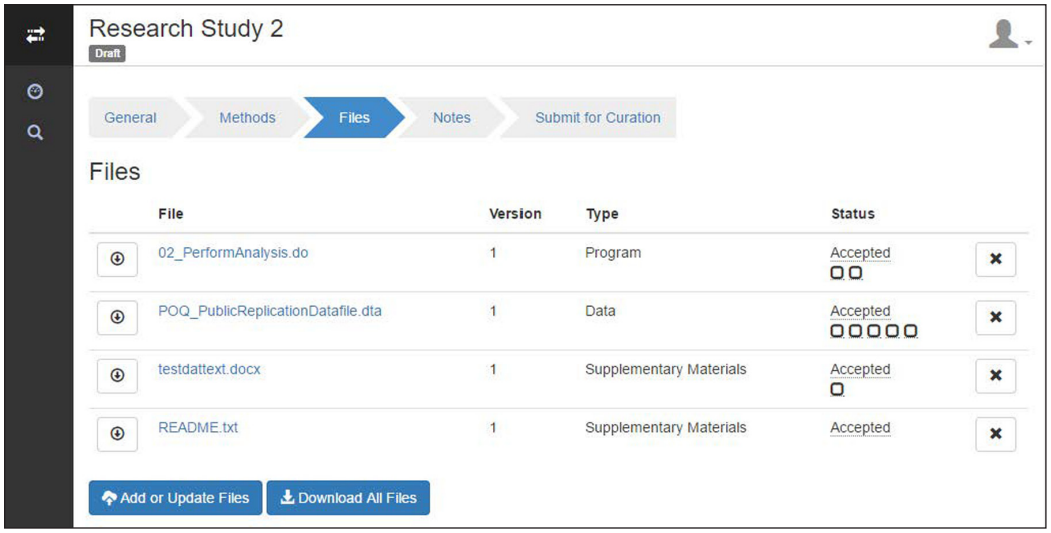

Figure 3: The Depositor view of the file list after initial upload, as seen in the user interface.

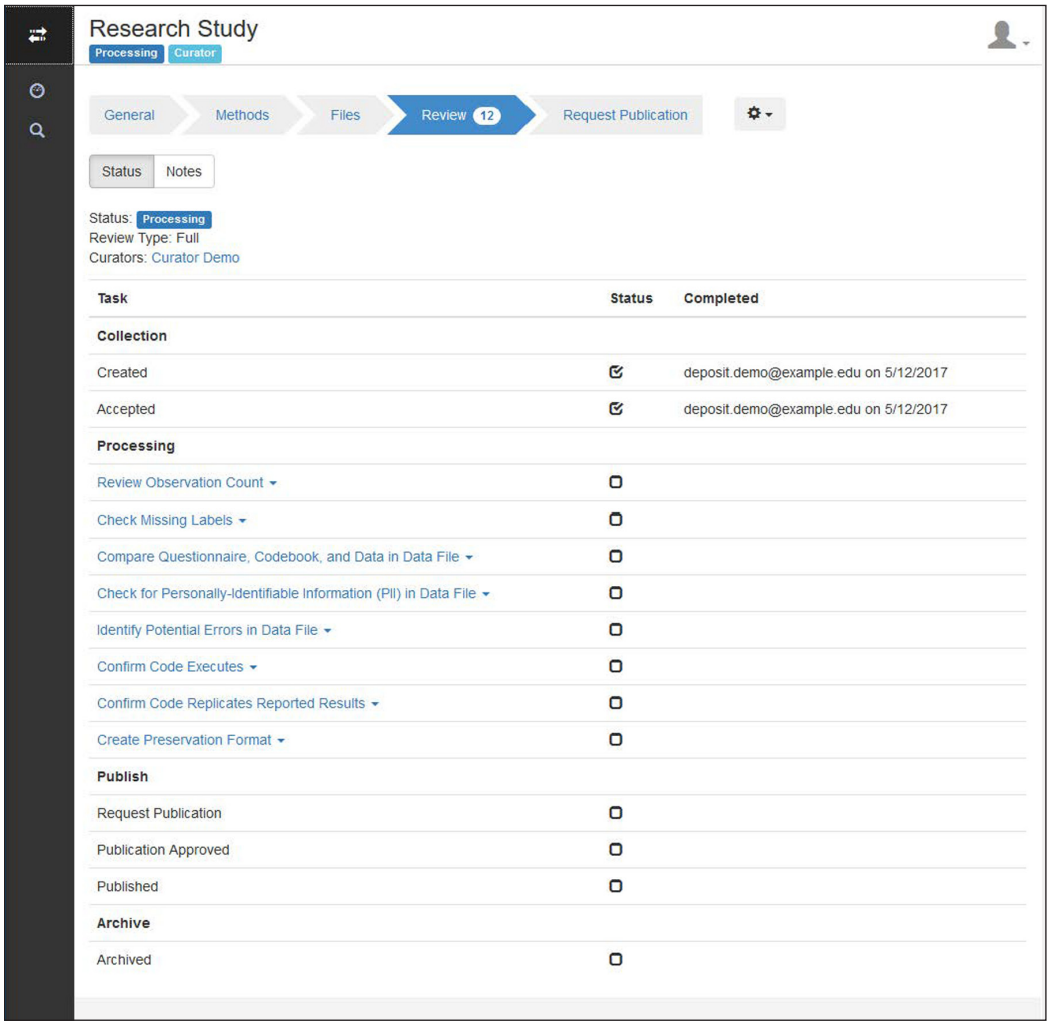

Figure 4: The Curator view of the curation tasks for the assigned record, as viewed in the user interface.

When notified of a new record, the Curator will complete all curation tasks which the tool automatically assigns based on the file types (for example, a data file will have different curation steps than a codebook). **Figure 4** is an example of the Curator's review panel which lists curation tasks. Once curation is complete, the Curator will submit the Catalog Record for publication approval by the Admin.

A record publication approval triggers a series of events on the server side. A plain-text, preservation copy (e.g., .csv) is created of certain proprietary data files (e.g., Stata .dta) and added to the 'processed' directory. Any updates or additions to the study-level metadata are synced in the Curation Database. Changes to the files are versioned by Git, which is built into the Curation Service. The Curation Database also stores the details about each completed curation step including the date, time, and which user completed the step. The tool provides a full history log for each Catalog Record. If configured, updates to the data files and variables are synced to the Metadata Database. The optional Colectica Repository software uses metadata from both the Curation & Metadata Databases to create a detailed metadata file using the Data Documentation Initiative (DDI) 3.2 schema. Finally, the Catalog Record, and each file marked as 'public', are assigned a persistent link. YARD uses Yale's in-house handle service to generate these links, but integration with other services is possible.

The Curation Workflow

The curation workflow as implemented in YARD is designed to the specifications of ISPS at Yale University. The workflow is based on the Inter-university Consortium for Political and Social Research (ICPSR, n.d.) pipeline and adapted for quantitative research output from randomized controlled trials (RCTs) in the social sciences (Peer et al, 2014). For example, YARD prompts Curators to review whether documentation and contextual information necessary for long-term usability (e.g., a codebook, a readme file) are included. The curation workflow has been further enhanced to include tasks for reviewing code and statistical analysis to obtain verification of computational reproducibility (Peer & Wykstra, 2015; Peer, 2017). For example, YARD guides Curators to review code files – statistical and other programming scripts – by verifying that the code executes and that the published scientific results can be computationally reproduced with the given code and data. The workflow was developed with input from potential users at the Odum Institute Data Archive at the University of North Carolina, Chapel Hill and the Cornell Institute for Social and Economic Research at Cornell University.

Integration with a Repository

The curation tool is not a replacement for a repository in so far as it is not meant to be an access point for other scholars or the general public. End users can only access records and files processed through the curation tool if they are ingested into another system such as a data repository or archive. The YARD implementation is currently designed to integrate with Drupal and provides access to processed records via the ISPS Data Archive.[7] Organizations can determine a preferred means of dissemination based on their own infrastructure.

There are two methods for integrating processed records into other software or workflows: the 'processed' directory and the Extensible Markup Language (XML) feed. The 'processed' directory is a compressed directory created by the curation tool when each catalog record is finalized and approved. The zipped archive, as seen expanded in **Figure 5**, is structured to match BagIt specifications.[8] It contains a file manifest with MD5 checksums, a handle map, and a 'data' directory containing the curated files and the application-generated DDI file. This DDI file contains the metadata necessary to ingest studies into another system or repository. Since studies can be re-curated and re-approved, a unique archive directory is created for each instance of publication (e.g., a study reviewed and finalized a second time will have two archive directories, one for each version).

The second option for disseminating records is the XML feed, which is created when a record is approved for publication. The feed contains metadata about each study and any associated files processed through the curation tool. **Figure 6** shows a sample of the XML feed. Only studies and files marked as 'public' will appear in the feed. The feed includes the persistent link for each file, so files can be downloaded or ingested via the feed. The feed can be ingested into any system with a XML mapping option. In the YARD implementation, the XML feed is ingested into a Drupal site using the Drupal feed importers module,[9] whereby each XML element is mapped to a specific Drupal field.

[7] See: https://isps.yale.edu/research/data (accessed 2019 November 18).
[8] See: http://www.dcc.ac.uk/resources/external/bagit-library (accessed 2019 November 27).
[9] See: https://www.drupal.org/project/feed_import (accessed 2019 November 18).

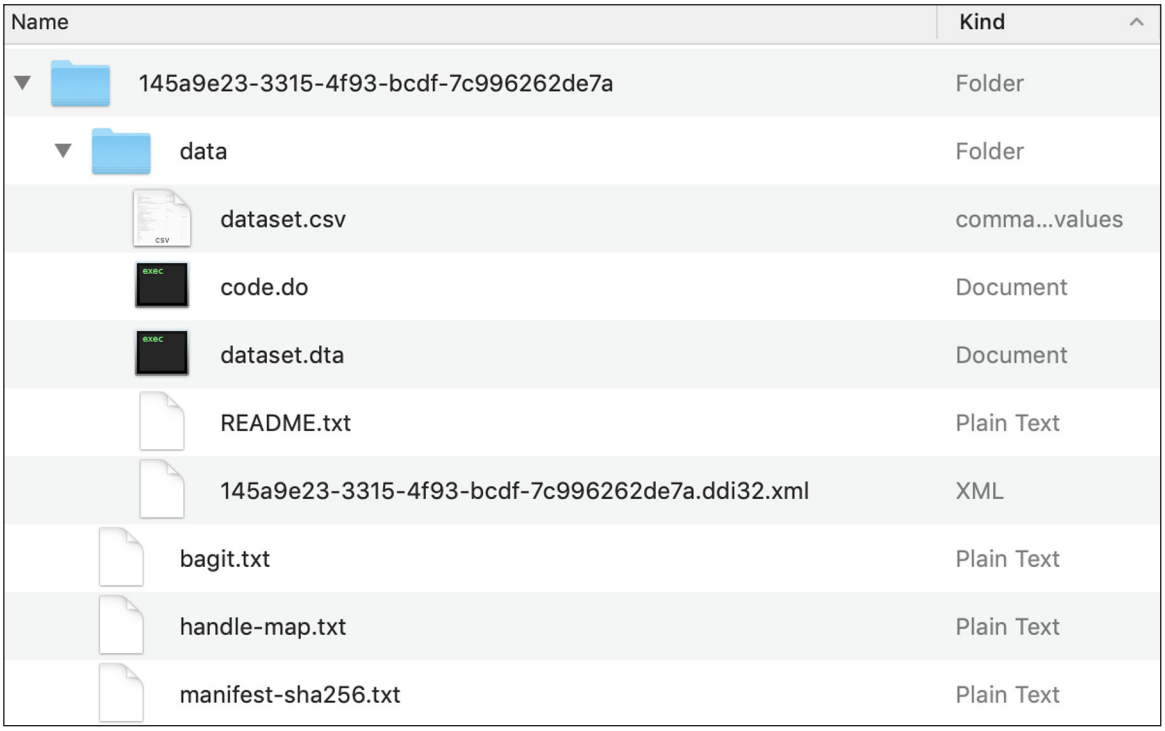

Figure 5: An example archive directory for a catalog record after curation and final approval.

```
<Record>
  <Guid>58aeb4e5-a3c6-4b4d-be82-7349ef1b9734</Guid>
  <Title>Title 2.16.2017</Title>
  <Author/>
  <Owner/>
  <Description>Description</Description>
  <StudyID>Number</StudyID>
  <StudyIDLower>number</StudyIDLower>
  <RelatedPublication>Related Pub</RelatedPublication>
  <RelatedProject>Related Proj</RelatedProject>
  <RelatedDatabase>Related DB</RelatedDatabase>
  <keywords>Keywords</keywords>
  <CreateDate>2017-02-16T21:10:58.867</CreateDate>
  <ResearchDesign>Field experiment</ResearchDesign>
  <DataType>Administrative (e.g., voting records) </DataType>
  <DataSource> Curation System</DataSource>
  <DataSourceInformation/>
  <CatalogRecordDataType/>
  <CatalogRecordDataSource/>
  <CatalogRecordDataSourceInformation/>
  <PersistentId/>
  <FieldDates>2016-12-06</FieldDates>
  <Location>N/A</Location>
  <LocationDetails>Location Details</LocationDetails>
  <UnitOfObservation>Individual</UnitOfObservation>
  <SampleSize>Sample Size</SampleSize>
  <InclusionExclusionCriteria>Inclusion Criteria
   </InclusionExclusionCriteria>
  <RandomizedProcedure>Random</RandomizedProcedure>
  <Treatment>Intervention</Treatment>
  <TreatmentAdministration>Door to door
   </TreatmentAdministration>
  <OutcomeMeasures>outcome mearures: unsure</OutcomeMeasures>
  <ArchiveDate>2017-02-16T21:34:38.757</ArchiveDate>
  <FileElement>
    <File>
      <id>13d4a6da-1bc6-4671-83a2-8c45e72ff0f8</id>
      <FileSize>27396</FileSize>
      <FileUrl>http://hdl.handle.net/10079/02v6x8n</FileUrl>
      <FileNumber>D777F556</FileNumber>
      <FileDescription>Data File</FileDescription>
      <FileFormat>.dta</FileFormat>
      <PublicFile>1</PublicFile>
      <CatalogRecordId>D777</CatalogRecordId>
    </File>
    <File>
      <id>64284add-a35b-49ad-a0b0-fdeda9501973</id>
      <FileSize>37376</FileSize>
      <FileUrl>http://hdl.handle.net/10079/v9s4n89</FileUrl>
      <FileNumber>D777F555</FileNumber>
      <FileDescription>Notes</FileDescription>
      <FileFormat>.doc</FileFormat>
      <PublicFile>1</PublicFile>
      <CatalogRecordId>D777</CatalogRecordId>
    </File>
    <File>
      <id>db8f0fc9-2fb3-42b7-9389-c25f6ad84aa7</id>
      <FileSize>35806</FileSize>
      <FileUrl>http://hdl.handle.net/10079/qjq2c7q</FileUrl>
      <FileNumber>D777F558</FileNumber>
      <FileDescription>Metadata (DDI 3.2)</FileDescription>
      <FileFormat>.xml</FileFormat>
      <PublicFile>1</PublicFile>
      <CatalogRecordId>D777</CatalogRecordId>
    </File>
  </FileElement>
</Record>
```

Figure 6: An example XML feed produced by YARD showing the metadata for a catalog record.

Customizability

The curation tool is designed to be fully customizable. That includes configuring the curation workflow such that curation tasks can be adjusted. For example, other curation frameworks could be applied (e.g., the Data Curation Network's CURATED checklist, see DCN, 2018) or specific tasks can be made optional or dropped altogether (e.g., checking for the presence of personally-identifiable information). These and other changes to the application code–changing user roles (e.g., the Depositor role may be eliminated if a repository only allows Curators to deposit, or other roles can be added), changing the metadata schema as appropriate to other disciplines, customizing study-level information (e.g., adding a geolocation field), and more–can be done by a skilled developer.

Other customization involves changes to admin settings in the application or editing config files on the web server. For example, within the application, administrators can customize file storage locations, edit or turn off persistent link minting, and integrate with other software to automate tasks where possible (see **Table 1**). Admin settings also allow customization of email notifications and permissions to create new users accounts. Finally, admin access to the web server grants permissions to edit config files in order to customize functions such as changing database paths, limit or increase allowable upload file size, and customize error reporting.

Given the diversity of research practices and products, the tool is designed to be modular and built with interchangeable components, such that it could be customized by an organization to meet its specifications, requirements, and policies.

Availability

YARD is currently supported at Yale with local IT resources and infrastructure, including admins who deploy the full stack and monitor and maintain the web and database servers. Any organization that assumes responsibility for the curation of research outputs–for example, a repository, a research lab, an academic research center, or a library–can have a local installation by compiling from the source code. Access to the code and comprehensive documentation are available in a public repository (Iverson & Smith, 2018). As an open source project, it is our hope that interested parties will join us in supporting and improving the software in accordance with best practice governance models in the academic Open Source community.

Discussion and Conclusions

This paper describes a policy-driven adaptable workflow tool supporting the archival and dissemination of high-quality research outputs. The tool is designed to increase the potential for long-term usability by creating high quality and FAIR-compliant data packages. The essential design principles applied to this tool is modularity and open source. The tool also promotes research transparency by connecting the activities of researchers, curators, and publishers through a single pipeline. Our vision is for this tool to be used by organizations committed to both rigorous research practices and high-quality output. We believe this project is a significant step toward the "development of more generic tools and processes for validating and improving various aspects of data quality," as called by Digital Curation Centre Director, Kevin Ashley (2013).

YARD addresses variability in research output quality by helping economize and standardize curation efforts and services. It achieves that by,

1. Providing a workflow in which curation activities can be managed, tracked, inspected, standardized, and shared and,
2. Enabling implementation of quality standards and policies aligned with making research outputs more usable and interpretable in the long term and deploying a design approach that facilitates accommodating new conditions and integrating with improved tools.

Developing YARD was the collaborative effort of several groups at Yale and Colectica. The team made use of project management tools to communicate with the developers, track software bugs, and document the software development process. At Yale, good working relationships with partners in Yale Information Technology Services and Yale University Library IT were essential to the project's success in all steps of development. Looking back at the trajectory of the project's development, we recognize that, as with many software development projects, we were subject to tightly resourced environments that presented a challenge to well-intentioned but sometime compromised efforts to test and deploy the tool within scheduled timelines and to assume local project ownership beyond the initial Colectica development. A more agile approach to deployment and testing of the software could have mitigated the consequences of some legacy decisions made at the project's inception, such as, hosting the software on Yale ITS managed infrastructure (which provided automated server backups and security management but required additional coordination across departments) as opposed to a cloud service like AWS (which would give us more flexibility and control but require additional internal resources). Despite a lack of funding beyond the initial development and unforeseen delays, we have confidence in YARD's sound fundamentals and potential to contribute to standardized, efficient, and transparent curation.

Future improvements to the software may include developing an API to allow further integration of published records with various workspaces or repository destinations. The curation log generated by the tool may be mined for information about curation tasks to inform staffing needs and educational efforts relating to research data management and curation. The curation tool's version tracking and UUID capabilities may be used to track the evolution of digital objects throughout the research lifecycle, from creation to publication or archival, and to link them to other systems, such as institutional sponsored projects record keeping. Related, other methods of authentication may be implemented to allow seamless integration with other systems. We urge the community to take advantage of the open source software. For now, we are confident that the curation tool provides a framework and a method for enhancing the digital artifacts underpinning scientific research – something that research institutions, repositories and archives, and publishers have a vested interest in.

Acknowledgements

We are grateful to our reviewers for helpful comments about this manuscript. We thank Yale University Library and Yale Information Technology Services for supporting this project. We give special thanks to Ann Green for contributing to the conceptualization of this project, to the team at Innovations for Poverty Action for early planning, to Themba Flowers for contributing to the deployment at Yale, and to Mike Friscia, Eric James, and Robert Wolfe for critical support at Yale. We also thank Jeremy Iverson and Dan Smith at Colectica for their dedicated development work, and Thu-Mai Christian at the Odum Institute Data Archive and Florio Arguillas at the Cornell Institute for Social and Economic Research for participating in early testing. This project would not have been possible without the support of the Institution for Social and Policy Studies at Yale University and its commitment to upholding and implementing the highest standards in academic research. L.P. acknowledges funding from Innovations for Poverty Action.

Competing Interests

The authors have no competing interests to declare.

References

Altman, M. 2012. Mitigating Threats to Data Quality Throughout the Curation Lifecycle. In: Marchionini, G, Lee, CA, Bowden, H and Lesk, M (eds.), *Curating for Quality: Ensuring Data Quality to Enable New Science*. Final Report: Invitational Workshop Sponsored by the National Science Foundation, Sept. 10–11, 2012, Arlington, VA. https://ils.unc.edu/callee/curating-for-quality.pdf.

Ashley, K. 2013. Data Quality and Curation. *Data Science Journal*, 12: GRDI65–GRDI68. DOI: https://doi.org/10.2481/dsj.GRDI-011

Blei, DM and **Smyth, P.** 2017. Science and Data Science. *Proceedings of the National Academy of Sciences*, 114(33): 8689–8692. DOI: https://doi.org/10.1073/pnas.1702076114

Burton, M, Lyon, L, Erdmann, C and **Tijerina, B.** 2018. *Shifting to Data Savvy: The Future of Data Science in Libraries*. Project Report. University of Pittsburgh, Pittsburgh, PA. http://d-scholarship.pitt.edu/id/eprint/33891.

Center for Research Libraries (CRL). 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf (accessed on 2020 April 2).

Chambers, K, et al. 2019. Towards Minimum Reporting Standards for Life Scientists. *MetaArXiv*. April 30. DOI: https://doi.org/10.31222/osf.io/9sm4x

Circle Systems, Inc. 2017. Stat/Transfer (Version 14). [computer software]. Available from https://stattransfer.com/.

Colectica. 2016. Colectica Repository (Version 5.0.4236). [computer software]. Available from https://www.colectica.com/software/repository/.

Data Curation Network. 2018. *Checklist of CURATED Steps Performed by the Data Curation Network*. http://z.umn.edu/curate. (also found at https://datacurationnetwork.org)

Digital Curation Centre (DCC). n.d. *What is Digital Curation?* http://www.dcc.ac.uk/digital-curation/what-digital-curation (accessed on 2019 December 30).

Dillo, I and **Leeuw, L.** 2018. CoreTrustSeal. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare*, 71(1): 162–170. DOI: https://doi.org/10.31263/voebm.v71i1.1981

International Organization for Standardization (ISO). 2012. *ISO 16363:2012 – Space Data and Information Transfer Systems – Audit and Certification of Trustworthy Digital Repositories*. Geneva:

International Organization for Standardization. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510 (accessed on 2020 April 2).

Inter-university Consortium for Political and Social Research (ICPSR). n.d. *Data Enhancement*. https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/ingest/enhance.html (accessed on 2020 April 2).

Iverson, J and **Smith, D.** 2020, January 8. Colectica/Curation: Initial Release (Version v0.9). Zenodo. DOI: http://doi.org/10.5281/zenodo.3600615

Johnston, LR, et al. 2018. How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication,* 6(General Issue): eP2198. DOI: https://doi.org/10.7710/2162-3309.2198

Leek, JT and **Jager, LR.** 2017. Is Most Published Research Really False? *Annual Review of Statistics and Its Application*, 4(1): 109–122. DOI: https://doi.org/10.1146/annurev-statistics-060116-054104

Miguel, E, et al. 2014. Promoting transparency in social science research. *Science*, 343(6166): 30–31. http://science.sciencemag.org/content/343/6166/30. DOI: https://doi.org/10.1126/science.1245317

National Academies of Sciences, Engineering, and Medicine. 2018. *Open Science by Design: Realizing a Vision for 21st Century Research*. Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/25116

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/25303

Peer, L. 2017. Enabling Scientific Reproducibility with Data Curation and Code Review. In: Johnston, L (ed.), *Curating Research Data Volume Two: A Handbook of Current Practice*. Chicago, Illinois: Association of College and Research Libraries. http://hdl.handle.net/11299/185335.

Peer, L, Green, A and **Stephenson, E.** 2014. Committing to Data Quality Review. *International Journal of Digital Curation*, 9(1): 263–291. DOI: https://doi.org/10.2218/ijdc.v9i1.317

Peer, L and **Wykstra, S.** 2015. New Curation Software: Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation. *IASSIST Quarterly*, 39(4): 6–13. https://iassistquarterly.com/index.php/iassist/article/view/902 (accessed on 2019 October 31). DOI: https://doi.org/10.29173/iq902

Sposito, FA. 2017. What Do Data Curators Care About? Data Quality, User Trust, and the Data Reuse Plan. *Paper presented at: IFLA 2017 Meeting*, Wrocław, Poland. http://library.ifla.org/id/eprint/1797.

Stodden, V, Seiler, J and **Ma, Z.** 2018. An Empirical Analysis of Journal Policy Effectiveness for Computational Reproducibility. *Proceedings of the National Academy of Sciences*, 115(11): 2584–2589. DOI: https://doi.org/10.1073/pnas.1708290115

The Royal Society. 2012. *Science as Open Enterprise*. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf (accessed on 2019 December 4).

Tiesi, G. 2014. ClamAV Native Win32 Port (Version 0.98.4). [computer software]. Available from https://github.com/clamwin/clamav-win32-old/tree/clamav-0.98.

Yale University. 2016. Yale Persistent Linking Service – Web Service (Version 1.0). [computer software]. Available from http://link.its.yale.edu/ypls-ws/PersistentLinking?wsdl.

How to cite this article: Peer, L and Dull, J. 2020. YARD: A Tool for Curating Research Outputs. *Data Science Journal,* 19: 28, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2020-028

Submitted: 08 December 2019 **Accepted:** 25 June 2020 **Published:** 15 July 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS