Reusable Data Best Practices Documentation

CD2H Data Working Group

Jan 12, 2021

Contents:

Access the Informatics Playbook GitHub repository to suggest changes, add content, or make comments.

CHAPTER 1

Chapter 1: Research Data Licensing

1.1 Intended audience

This guidance is primarily targeted to providers of publicly-disseminated research data and knowledge and to the funders thereof. Many licensing possibilities for a data resource are taken into account; however, in a some cases the point-of-view is focused from one direction, which can reduce the clarity of our curations for the informatics community. In these cases, we may take on the role of a noncommercial academic group that is based in the US and creating an aggregating resource, noting that other entities may have different results in the license commentary.

1.2 Why is this important?

The increasing volume and variety of biomedical data have created new opportunities to integrate data for novel analytics and discovery. Despite a number of clinical success stories that rely on data integration (rare disease diagnostics, cancer therapeutic discovery, drug repurposing, etc.), within the academic research community, data reuse is not typically promoted. In fact, data reuse is often considered not innovative in funding proposals, and has even come under attack (the now infamous Research Parasites NEJM article).

The FAIR principles–Findable, Accessible, Interoperable, and Reusable–represent an optimal set of goals to strive for in our data sharing, but they do little to detail how to actually realize effective data reuse. If we are to foster innovation from our collective data resources, we must look to pioneers in data harmonization for insight into the specific advantages and challenges in data reuse at scale. Current data licensing practices for most public data resources severely hamper reuse of data, especially at scale. Integrative platforms such as the Monarch Initiative, the NCATS Data Translator, the Gabriella Miller Kids First DCC, and the myriad of other cloud data platforms will be able to accelerate scientific progress more effectively if these licensing issues can be resolved. As affilated with these various consortia, Center for Data to Health (CD2H) leadership strives to facilitate the legal use and reuse of increasingly interconnected, derived, and reprocessed data. The community has previously raised this concern in a letter to the NIH.

How reusable are most data resources? In our recently published manuscript, we created a rubric for evaluating the reusability of a data resource from the licensing standpoint. We applied this rubric to over 50 biomedical data and knowledge resources. Custom licenses constituted the largest single class of licenses found in these data resources. This suggests that the resource providers either did not know about standard licenses or felt that the standard licenses

did not meet their needs. Moreover, while the majority of custom licenses were restrictive, just over two-thirds of the standard licenses were permissive, leading us to wonder if some needs and intentions are not being met by the existing set of standard permissive licenses. In addition, about 15% of resources had either missing or inconsistent licensing. This ambiguity and lack of clear intent requires clarification and possibly legal counsel.

Putting this all together, a majority of resources would not meet basic criteria for legal frictionless use for downstream data integration and redistribution activities despite the fact that most of these resources are publicly funded, which should mean the content is freely available for reuse by the public.

1.3 Takeaways

To receive a perfect reusability score, the following criteria should be met:

1.3.1 A) License is public, discoverable, and standard

1.3.2 B) License requires no further negotiation and its scope is both unambiguous and covers all of the data

1.3.3 C) Data covered by the license are easily accessible

1.3.4 D) License has little or no restrictions on the type of (re)use

1.3.5 E) License has little or no restrictions on who can (re)use the data

The full rubric is available at http://reusabledata.org/criteria.html

1.3.6 Lessons learned:

The hardest data to license (in or out) are often data integrated from multiple sources with missing, heterogeneous, nonstandard, and/or incompatible licenses. The opportunity exists to improve this from the groud up. While the situation will never be perfect, it could be substantially improved with modest effort.

1.4 Acknowledgments

ReusableData.org is funded by the National Center for Advancing Translational Sciences (NCATS) OT3 TR002019 as part of the Biomedical Data Translator project. The (Re)usable Data Project would like to acknowledge the assistance of many more people than can be listed here. Please visit the about page for the full list.

CHAPTER 2

Chapter 2: Identifier best practices

2.1 Intended audience

We propose actions that identifier practitioners (public database providers) should take in the design, provision, and reuse of identifiers. We also outline important considerations for those referencing identifiers in various circumstances, including by authors and data generators. While the importance and relevance of each lesson will vary by context, there is a need for increased awareness about how to avoid and manage common identifier problems, especially those related to persistence and web-accessibility/resolvability. We focus strongly on web-based identifiers in the life sciences; however, the principles are broadly relevant to other disciplines. Although the lessons are most relevant to publicly-accessible research data, there are transferrable principles for private data as well.

2.2 Why is this important?

The issue is as old as scholarship itself: readers have always required persistent identifiers in order to efficiently and reliably retrieve cited works. "Desultory citation practices" have been thwarting scholarship for millennia either because reliable identifiers were unavailable or because authors failed to use them. While the internet has revolutionized the efficiency of retrieving sources, the same cannot be said for reliability; it is well established that a significant percentage of cited web addresses go "dead". This process is commonly referred to as "link rot" because availability of cited works decays with time. Although link rot threatens to erode the utility and reproducibility of scholarship, it is not inevitable; link persistence has been the recognized solution since the dawn of the internet. However, this problem, as we will discuss, is not at all limited to referencing journal articles. The life sciences have changed a lot over the past decade as data have evolved to be ever larger, more distributed, more interdependent, and more natively web-based. This transformation has fundamentally altered what it even means to "reference" a resource; it has diversified both the actors doing the referencing and the entities being referenced. Moreover, the challenges are compounded by a lack of shared terminology about what an "identifier" even is.

2.3 Status and contribution mechanisms

This chapter is implementation-ready. We welcome feedback, whether by way of github issue, google form, or email us.

2.4 Takeaways

The list of lessons is below; the paper from which they are derived contains examples and rationale for each one.

2.4.1 Lesson 1. Credit any derived content using its original identifier

2.4.2 Lesson 2. Help local IDs travel well; document prefix and patterns

2.4.3 Lesson 3. Opt for simple, durable web resolution

2.4.4 Lesson 4. Avoid embedding meaning or relying on it for uniqueness

2.4.5 Lesson 5. Design new identifiers for diverse uses by others

2.4.6 Lesson 6. Implement a version-management policy

2.4.7 Lesson 7. Do not reassign or delete identifiers

2.4.8 Lesson 8. Make URIs clear and findable

2.4.9 Lesson 9. Document the identifiers you issue and use

2.4.10 Lesson 10. Reference and display responsibly

Better identifier design, provisioning, documentation, and referencing can address many of the identifier problems encountered in the life science data cycle, leading to more efficient and effective science. However, best practices will not be adopted on the basis of their community benefit alone; the practices must be both easy and rewarding to the groups that do the implementing. In the broader context of scholarly publishing, this is just what DOIs afford; DOIs succeeded because they were well aligned with journals' business goals (tracking citations) and because the cost was worth it to them. However, in the current world where everyone is a data provider, alignment with business goals is still being explored. Meta resolvers can provide a use case for journals and websites seeking easier access to content, while software applications leverage these identifier links to mine for knowledge.

We recognize that improvements to the quality, diversity, and uptake of identifier tooling would lower barriers to adoption of the lessons presented here. Those that issue data identifiers face different challenges than do those referencing data identifiers. We understand there are ecosystem-wide challenges and we will address these gaps in the relevant initiatives. We also recognize the need for formal software-engineering specifications of identifier formats and/or alignment between existing specifications. Here, we implore all participants in the scholarly ecosystem—authors, data creators, data integrators, publishers, software developers, resolvers—to aid in the dream of identifier harmony and hope that this playbook can catalyze such efforts.

2.5 Acknowledgments

The content of this chapter was derived from the following paper, an open access article distributed under the terms of the Creative Commons Attribution License that permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Numerous funding sources were involved in supporting this effort, most notably, BioMedBridges and the Monarch Initiative; however, all of the sources are listed in the paper.

- Julie A. McMurry, Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot, John Deck, Michel Dumontier, Donal K. Fellows, Alejandra Gonzalez-Beltran, Philipp Gormanns, Jeffrey Grethe, Janna Hastings, Jean-Karim Hériché, Henning Hermjakob, Jon C. Ison, Rafael C. Jimenez, Simon Jupp, John Kunze, Camille Laibe, Nicolas Le Novère, James Malone, Maria Jesus Martin, Johanna R. McEntyre, Chris Morris, Juha Muilu, Wolfgang Müller, Philippe Rocca-Serra, Susanna-Assunta Sansone, Murat Sariyar, Jacky L. Snoep, Stian Soiland-Reyes, Natalie J. Stanford, Neil Swainston, Nicole Washington, Alan R. Williams, Sarala M. Wimalaratne, Lilly M. Winfree, Katherine Wolstencroft, Carole Goble, Christopher J. Mungall, Melissa A. Haendel, Helen Parkinson. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Bio.* 2017. https://doi.org/10.1371/journal.pbio.2001414

CHAPTER 3

Chapter 3: Sharing Educational Resources

3.1 Intended audience

Educators at CTSAs who want to build a community of learners around their educational material, or make their educational materials discoverable and usable by others.

3.2 Current version / status

3.3 Guidance

Actively Soliciting Comments and Feedback

3.4 Lessons learned / summary

- An effective method of sharing your materials across CTSAs is through the CLIC Education Clearinghouse
- Making educational resources *discoverable* requires publishing metadata about that resource
- Consider making your resource an open educational resource (OER) by sharing the material openly
- If your resource is an OER, make extra metadata available that emphasizes its reuse
- Inviting other instructors and learners to utilize and update educational resources builds an educational community
- Educational communities have multiple benefits beyond keeping course material up to date, including providing support for both beginner and intermediate learners

3.5 Why this is important

Keeping educational resources both *discoverable* and *up to date* is difficult for single CTSA sites.

3.5.1 Discoverability

Making an educational resource *discoverable* is of vital importance. Sites replicating educational material that already exists is ultimately a poor use of resources. Such replication largely ignores the possibility of building educational communities around such resources.

Discoverability may be done for multiple reasons:

1. to advertise that an educational resource exists so others may take the course,
2. whether a course is appropriate to *reuse*, *repurpose*, or *remix*, and has additional resources to aid other instructors in adopting the material.

3.5.2 Keeping Material Updated

One alternative to a single site maintaining an education resource is providing *collective ownership* to a learning community, such as the model for The Carpentries (Software, Data, and Library Carpentry), or the R for Data Science learning community. Hundreds, if not thousands of educators have tested, honed, and improved the lesson material for these groups. Collective ownership of the learning material makes the material stronger and more applicable to a wide range of learners. Additionally, the learning community that results from such collective ownership provides an opportunity for those with intermediate skills to improve their knowledge and practice this knowledge in a safe and supportive environment.

3.6 Status and feedback mechanisms

What stage is it in the development process? (See options below). Description and links for how people can comment, or contribute – whether via Google form or GitHub issue etc.

3.7 Takeaway List

1. Submit your educational resource to the CLIC Educational Clearinghouse
2. Consider making your resource an Open Educational Resource (OER)
3. Make metadata available for an educational resource available by publishing metadata using a standard such as *MIER* or *Harper Lite*
4. Add metadata that encourages reuse of your educational resource
5. Encourage the formation of an educational community around an educational resource
6. Foster growth and updates of your material through quarterly hackathons or sprints

3.8 Deep dive into takeaways

3.8.1 1. Submit your educational resource to the CLIC Educational Clearinghouse

3.8.2 2. Consider making your resource an Open Educational Resource

Making your educational resource open has many benefits.

3.8.3 3. Make metadata available for an educational resource by publishing metadata using a standard such as *MIER* or *Harper Lite*

At the very least, map your educational resource to the Clinical and Translational Science Competencies established by CLIC. Follow the trend in tags (keywords) that are commonly used.

In order for your resource to be discoverable, providing essential metadata using a standard such as *MIER* or *Harper Lite* is important.

3.8.4 4. Add metadata that encourages reuse of your educational resource

Both the MIER and Harper-lite metadata standards include metadata that are specific to reusing course material:

1. What is the Licensing? Is the resource available to be repurposed by others?

For many instructors, if the licensing is too restrictive (such as requiring the No-Derivatives), instructors may be prevented from reusing materials. Consider licenses such as CC-BY-NC (Non Commercial), which is permissive for those who use the material for Non-Commercial uses.

1. Who is the audience? Who is the material for?
2. Are instructor notes available?

For their workshops, The Carpentries include extensive instructor notes that cover what was and was not successful during a workshop; such a resource is invaluable to understanding whether the material is written at an appropriate level for learners.

1. Is there a code of conduct?

3.8.5 5. Encourage the formation of an educational community around an educational resource

A quick and simple way to encourage community formation is to start a Slack channel associated with a resource. Encourage discussion and questions there.

Be responsive to feedback and be willing to give contributor roles to people who suggest changes to the material.

3.8.6 6. Foster growth and updates of your material through quarterly hackathons or sprints

3.9 Acknowledgments

CHAPTER 4

Chapter 4: Managing Translational Informatics Projects

4.1 Intended audience

Managers of Translational Informatics Projects

4.2 Why is this important

Translational Informatics projects are increasingly cross-institutional and even international; however, managing them comes with many shared pain points. This guidance will help anyone who is organizing or managing cross-functional distributed teams that develop code or that analyze data across the translational divide. Specifically, we will introduce several practical tools and techniques for managers to facilitate these kinds of endeavors. Exercises in the companion tutorial will familiarize participants with helpful tools and techniques, and help them make informed decisions about which tools might work best for their particular contexts. We conclude with a session wherein all participants are welcome to share additional pain points and related experience.

4.3 Takeaways

4.3.1 Pick the project management technique that is appropriate for your project (Agile, Waterfall Model, Kanban, etc)

4.3.2 Understand the implications of that management technique for the full lifecycle of your project

4.3.3 Get familiar with your (diverse) stakeholders

4.3.4 Have a process for team onboarding (Some guidance here for using forms)

4.3.5 Have a process communications

4.3.6 Have a process for shared document management

4.3.7 Organize work into a roadmap that is clear and achievable

4.3.8 Pick the work tracking platform that is right for your (whole) team

4.3.9 Focus your planning efforts in discrete horizons (eg. 2 weeks, 2 months, 2 years)

4.3.10 Make sure that all of the work that is planned is also assigned

4.3.11 Don't make security or licensing an afterthought

4.3.12 Don't be afraid to course correct

4.4 Acknowledgments

Thanks to Justin Ramsdill for the Agile introduction (also linked from the MTIP tutorial site).

CHAPTER 5

Chapter 5: Software and Cloud Architecture

Reusable Data Best Practices Guidebook

Chapter 5: Software and Cloud Architecture (draft, v1.0)

1. **Intended audience(s):A. CTSA hub leaders** (strategic recommendations for the use of cloud computing and reusable software resources at the hub and network levels)**B. Clinical and translational scientists** (project-level recommendations for the use of cloud computing and reusable software resources to meet individual needs and enhance the reproducibility, rigor, and shareability of research products)**C. Informatics and technology solution providers** (technical recommendation for how to access and use CD2H provisioned cloud computing resources and reusable software components)
2. **Current version / status:A. Last revision:** 12/18/2019**B. Status:** draft, outline
3. **Lessons learned / summary:A. Mission and purpose of the CD2H Tool and Cloud Community Core:** Computational technologies and tools are vital to clinical and translational research; however, CTSA hubs currently develop, deploy, and manage these key resources independently. As a result, these processes are tedious, costly, and heterogeneous. This core will address these issues by establishing a common tool and cloud computing architecture, and will provide CTSA hubs with an affordable, easy to use, and scalable deployment paradigm. Such an approach will support a robust ecosystem that demonstrates the use of shared tools and platforms for the collaborative analysis of clinical data. Hubs can easily promote and deploy their own products as well as adopt others, thereby transcending long-standing "boundaries" and solving common and recurring information needs.**B. Value and vision:C. Dimensions of tool and cloud architecture and capabilities:**

   i. **Cloud hosting** for software applications and platforms, leveraging Amazon Web Services (AWS) environment managed by NCATS and provisioned by CD2Hii. **Tool registry** to assist in the sharing and quality assurance of shared software components developed by CTSA hubs **<LINK TO SLIDES RE: TOOL REGISTRY PROJECT>**iii. **Build and test framework** for collaborative software development projectsiv. **Sandboxes** to provide spaces for informatics-focused workgroups seeking solutions to shared data analytic and management challengesv. **Benchmarking** of algorithms and predictive models using Challenge framework
4. **Status and feedback mechanisms:**A. **CD2H cloud hosting architecture** (v1.0) currently available for community feedback and comments:
   1. CD2H-NCATS Cloud Architecture proposal

    2. Architecture Response FormB. **CD2H cloud resource request "intake" form** (process for requesting access to CD2H provisioned cloud infrastructure)i. Cloud resource request intake formii. Cloud deployment projects dashboard (under development)C. **Prototype shared tools** deployed using NCATS/CD2H cloud resources or other Tool and Cloud Community Core capabilities:i. Competitions (peer review and competitive application management)ii. Leaf (platform agnostic clinical data browser)D. Program-wide **CD2H tool registry**i. CD2H LabsE. **Benchmarking projects** leverage Challenge framework:i. Metadata Challenge (sharing of cancer-focused datasets)ii. EHR Challenge (mortality prediction)

5. **Takeaway list:**A. Create a common cloud computing architecture that can enable the rapid deployment and sharing of reusable software components by CTSA hubsB. Demonstrate the use of shared tools and platforms for the collaborative analysis of clinical data in a manner that transcends individual CTSA hub "boundaries"C. Disseminate a common set of tools that can be employed for both the local and collaborative query of common data warehousing platforms and underlying data modelsD. Pilot the "cloudification" of software artifacts that can be shared across CTSA hubs to address common and recurring information needs.
6. **Deep dive into takeaways:**A. CD2H-NCATS Cloud Deployment ChecklistB. CD2H-NCATS Cloud Deployment Process WorkflowC. CD2H-NCATS Architecture Design ProposalD. CD2H-NCATS Architecture Request for Feedback FormE. CD2H-NCATS Federated Authentication (UNA) OverviewF. Code and documentation repositories for ongoing Tool and Cloud Community Core projects:
    1. Tool-Cloud-Infrastructure Core GitHub repo
    2. Cloud-Tool-Architecture project GitHub repo
    3. Competitions project GitHub repo
    4. EHR Dream Challenge project GitHub repo
7. **Acknowledgements**
8. <LIST CLOUD CORE PARTICIPANTS>

5.1 Cloud Collaboration software

5.2 Cloud architecture

5.3 Software best practices

CHAPTER 6

Chapter 6: Understanding Data Harmonization

6.1 Purpose & intended audience

This resource offers guidance to members of the CTSA informatics community including information about Data Harmonization that key stakeholders (leadership, researchers, clinicians, CIOs) can use at their institutions. This guidance can be useful to those who are new to Data Harmonization, as well as to those who are experts and may need assistance conveying the importance of Data Harmonization to a lay audience.

6.2 Why is this important?

Clinical data are among the most valuable artifacts within CTSA hubs. Appropriately leveraging these data for translational research purposes, while respecting privacy and honoring hub autonomy, will advance CTSA goals and demonstrate its power as a network. The Health Level 7 (HL7) FHIR standard has the potential to enable hubs to develop a next-generation repository from application program interfaces (APIs) that are built into every electronic health record (EHR). For optimal harmonization, these APIs need to be integration-ready, whether used directly for federated queries or for transformation to any number of common standards.

6.3 Status and how to contribute:

This document is currently in version 1.0 release. Comments from the community are appreciated for incorporation of next version. A commentable copy in Google doc form is available.

The live version (Chapter 6, Version 1.0) is rendered at:

https://reusable-data-best-practices.readthedocs.io

6.4 Takeaways

The categories of best practice include the following

- Data Harmonization Mission
- Governance
- Sustainability
- Workforce
- Infrastructure
- Relationship w the clinical enterprise
- Data practices
- External relationships and outreach

6.5 Acknowledgments

Co-leads: Boyd Knosp, University of Iowa (https://orcid.org/0000-0002-3834-3135); Catherine K. Craven, Icahn School of Medicine at Mount Sinai

Christopher G. Chute, Johns Hopkins University (https://orchid.org/0000-0001-5437-2545); Jeanne Holden-Wiltse, University of Rochester CTSI (https://orcid.org/0000-0003-2694-7465); Laura Paglione, Spherical Cow Group (https://orcid.org/0000-0003-3188-6273); Svetlana Rojevsky, Tufts Clinical and Translational Science Institute (https://orcid.org/0000-0002-8353-9006); Juliane Schneider, Harvard Catalyst | Clinical and Translational Science Center (https://orcid.org/0000-0002-7664-3331); Adam Wilcox, University of Washington.

Edited by:

- Lisa O'Keefe | Northwestern University | 0000-0003-1211-7583 | CRO:0000065
- Charisse Madlock-Brown | University of Tennessee Health Science Center | 000-0002-3647-1045
- Andréa Volz | Oregon Health & Science University | 0000-0002-1438-5664

6.6 Funding:

This work was supported by the National Institutes of Health's National Center for Advancing Translational Sciences CTSA Program Center for Data to Health (Grant U24TR002306).

CHAPTER 7

Chapter 7: Repository architecture and culture to support research at the local CTSA hub level

7.1 Intended Audience

This guidance is intended for academic institutional repository stakeholders, such as: (1) researchers with various goals: finding collaborators, seeking datasets for secondary analysis and teaching, sharing data for compliance with journal and funder mandates, having themselves and their team be recognized for the entirety of their research output (and make that output citable); (2) undergraduate and graduate students for learning and research; (3) support staff who assist researchers in carrying out tasks; and (4) departments, cores, institutions, and consortia who share, aggregate and report data. Next generation institutional repository architecture can meet the needs of all the aforementioned stakeholders and beyond. An example can be seen in the development of InvenioRDM, a turnkey research data management repository.

7.2 Why is this important?

In recent years, expansion of the institution's role in managing research outputs has been associated with increased scientific reproducibility; open science; enhancement of discovery, access, and use of information; increased collaboration and interdisciplinary research; and increased long-term stewardship of scholarly outputs, according to the MIT Report on the Future of Libraries, 2016. This management is frequently accomplished through an "open, trusted, durable, interdisciplinary, and interoperable content platform" that supports research data throughout its lifecycle. The Confederation of Open Access Repositories (COAR), an organization serving repositories and repository networks around the globe, released guiding principles for these platforms in 2018. The recommendations include:

- Distribution of control: distributed control of scholarly outputs, including preprints, post-prints, research data, and supporting software
- Inclusiveness and diversity: allowing for the unique needs of different regions, disciplines and countries
- Public good: technologies, protocols, and architectures of content repositories are available to everyone
- Intelligent openness and accessibility: scholarly resources to be made openly available
- Sustainability: institutional work toward long-term sustainability of resources

- Interoperability: content repositories' employment of functionalities and standards to ensure interoperability across the Web

While institutional repositories can serve as tools employing the six guiding principles, next generation repositories are increasingly enabling far greater interoperability, through both their frameworks and the data and metadata standards they employ, as well as increased support for sustainability and intelligent openness.

In the United States, open science and FAIR data practices have been increasingly supported by the scientific community since the 2013 release of a memo from the White House Office of Science and Technology Policy (OSTP) requiring federal agencies with over $100 million in annual conduct of research expenditures to develop plans to increase public access to the results of that research, including research data. While some agencies have set clear guidelines for the repositories to use for housing federally-funded research data, others have left the choice up to the researcher. The NIH Data Management and Sharing Policy finalized in Oct 2020 with an effective date of Jan 25, 2023 also leaves the decision up to the researcher as to where to deposit research data for sharing. This policy clearly states the critical role of research data management (RDM) and data sharing:

"Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data."

With no shortage of data repositories, researchers require guidance on where their data can best be deposited to maximize access, ensure compliance with journal and funder data sharing requirements, and ensure security and long-term preservation. COAR has released a Best Practices Framework for Repositories with behaviors that further reflect the needs of researchers when depositing, sharing, and storing data:

- Discovery
- Access
- Reuse
- Integrity and Authenticity
- Quality Assurance
- Privacy of Sensitive Data (e.g. human subjects)
- Preservation
- Sustainability and Governance
- Other Characteristics

These desired behaviors of next generation repositories reflect researchers' needs to make their research inclusive, participatory, and reproducible. These functions can be enabled through increased interaction with and interoperability of resources; support for commentary and annotation; support for discovery through navigation, user identification, verification, profiles, and alerts; and integration with other systems for managing research output and measuring research impact such as the Current Research Information System (CRIS) systems.

7.3 Takeaway List

The architecture of a research repository that is able to support emerging research needs meets all the specifications of the next generation repository as outlined above, and is modular and scalable, allowing for improvements in the future based on evolving user needs. InvenioRDM is an exemplar next generation repository, and its modular architecture and strong use of standards helps ensure the ability of the platform to support best practices in research data management. Each record in InvenioRDM is minted a DOI, a permanent identifier exposed through DataCite that is available for citation and compliance with data sharing requirements. Robust metadata, an open API, and the powerful Elasticsearch full-text search engine ensures that deposited data is findable, accessible, interoperable, and reusable (FAIR), and also allows for discovery through navigation and batch discovery of resources. As part of the "Reusable" element of FAIR data, licenses are declared with records in InvenioRDM to make the terms of reuse immediately clear. These features

of the InvenioRDM architecture support data sharing, innovation, knowledge dissemination, and interdisciplinary collaboration.

Users' needs are at the heart of the repository architecture of InvenioRDM, and to that end we are implementing specified controls and permissions that allow for identification and authentication of users, including support for ORCiD identifiers. InvenioRDM has an open API that makes it easy to share data with external resources, such as CRIS systems. InvenioRDM will provide users with the ability to create Collections, Communities, and shared private records, and will include social features. For ease of use, resource transfer is set up to allow a user to download resources in the same format in which they were uploaded. Industry standard usage statistics are collected for all record pages, including altmetrics, and tracking adheres to General Data Privacy Regulation (GDPR). Finally, the InvenioRDM architecture adheres to the Open Archival Information System (OAIS) standard and allows e.g. the retention of previous versions of records and a robust back-end database employing checksums and fixity checks to ensure long-term preservation of deposited digital files.

To support local RDM, institutions can foster a culture of research data management training, support, and best practices. Resources such as this playbook and guidance provided through informational sessions on responsible conduct of research and data management, data consultations, and support for using a repository solution like InvenioRDM, provided in a systematic way by data-focused professionals, will help researchers manage data throughout the research data lifecycle, from project conception through data collection, processing and analysis, dissemination, and preservation. It is important to emphasize that a repository like InvenioRDM can play a key role in each stage of the data lifecycle by serving as a place to find datasets for preliminary or feasibility studies, a place for researchers to find collaborators for the life of a project, and a place to safely disseminate and preserve data.

To reap the greatest benefits from the next generation repository features of InvenioRDM, create robust records that make the most of their many features, consider these **Top 5 Rules for Depositing Research Object Records**:

1. Make your deposit open access if possible
2. Use the appropriate license, see Reusable Data Best Practices Chapter 1
3. Add meaningful metadata to records
4. Attribute credit where credit is due (attribution chapter link)
5. Make sure you do not include any personal identifiable information (PII) in the record

7.4 Status and Feedback Mechanisms

The next generation repository InvenioRDM was launched with an alpha version at the end of October 2019. The Product Roadmap Overview can be seen here, and the Invenio Project Board, outlining future month-long project sprints, can be seen here. The InvenioRDM team also maintains a public GitHub site where Issues can be added regarding metadata, user interface requirements, and more.

Daily updates are available on a public Gitter chat. Monthly updates are made at the Resource Discovery Core meetings (open to the CD2H community) typically held on the last Thursday of the month at 1:00pm ET. The rolling meeting notes can be seen here. To contact the InvenioRDM team, please use the CD2H #InvenioRDM Slack channel.

7.5 Current Version

InvenioRDM enables organizations to securely house research products and make them discoverable, shareable, and citable, from publications and datasets to training materials, software, study materials, lay summaries, policy documents, and more. The platform is being developed as part of a large, multi-organization collaboration which includes the Center for Data to Health (CD2H) in partnership with the European Organization for Nuclear Research (CERN),

along with fourteen additional project partners. It is currently in the alpha release stage, with an example instance customized for Northwestern University acting as a showcase since October 2019. Another instance for demonstration purposes will be released at CERN in 2020.

7.6 Contributors to this guidebook chapter

Name | site | ORCID

- Sara Gonzales | Northwestern University | 0000-0002-1193-2298
- Lisa O'Keefe | Northwestern University | 0000-0003-1211-7583
- Guillaume Viger | Northwestern University |
- Matt Carson | Northwestern University | 0000-0003-4105-9220
- Tom Morrell | Caltech Library | 0000-0001-9266-5146
- Carlos Fernando Gamboa | Brookhaven National Laboratory
- Lars Holm Nielsen | CERN |
- Kai Wörner | Universität Hamburg | 0000-0001-8939-4437
- Kristi Holmes | Northwestern University | 0000-0001-8420-5254
- Andréa Volz | Oregon Health & Science University | 0000-0002-1438-5664

7.7 Acknowledgments

- Brookhaven National Laboratory
- Caltech Library
- CERN
- Data Futures
- Helmoltz Zentrum Dresden Rossendorf (HZDR)
- National Center for Data to Health (CD2H)
- Northwestern University Feinberg School of Medicine and Galter Health Sciences Library, DIWG & DIWG Metadata Subcommittee
- OpenAIRE
- TIND
- Tübitak Ulakbim
- TU Graz
- Unversität Hamburg
- WWU Münster

7.8 Funding:

This work was supported in part by the CERN Knowledge Transfer Fund, the National Institutes of Health's National Center for Advancing Translational Sciences CTSA Program Center for Data to Health (Grant U24TR002306), and through the many contributions of the project partners listed at the InvenioRDM project website.

CHAPTER 8

Chapter 8: Best practices for attribution and use of attribution

8.1 Intended Audience

Individuals include scholars working in academic and non-academic institutions, libraries, industry, etc. Groups include but are not limited to university administrators, and funding agencies.

8.2 Current Version

This draft is part of the Reusable Data Best Practices Playbook as a new chapter. Feedback is still be actively solicited and welcomed, given the "living" nature of this communication mode.

8.3 Why is this important?

It is very difficult to know who is contributing to research and what those contributions are. There has been a fundamental shift to recognize both the interdisciplinary, team-based approach to science, as well as the hundreds and thousands of more fine-grained contributions of varying types and intensities that are necessary to move science forward. Unfortunately, little infrastructure exists to identify, aggregate, present, and (ultimately) assess the impact of these contributions. These significant problems are technical as well as social and require an approach that assimilates cultural and social aspects of these problems in an open and community-driven manner. Ongoing efforts include the development of a contribution role ontology (built on CRedIT through the CRedIT ontology) to support modeling of the significant ways in which the translational workforce contributes to research.

Tracking and providing attribution for diverse contributions across the workforce support giving credit for work, allowing for a better understanding of what skills and activities are needed, and incentivizing participation in research. Moreover, this work helps to support and enhance a collaborative informatics community by fostering and promoting the development of an academic attribution and reimbursement framework for informatics products and processes. These processes could also help facilitate a contribution role that can be used for academic promotion and recognition.

8.4 Status

The Contributor Attribution Model is currently under development here. The Contributor Role Ontology (CRO) is released and available for use, with another release before the end of 2019. More information on the CRO is available here.

8.5 Feedback

- Architecting Attribution Engagement Page - Provides details on areas where the team is looking for help, how to contribute. This page also shares information about events and provides a call to participants to contribute ideas here, too.
- CD2H #Attribution Slack Channel - Project specific channel on Slack's Instant messaging platform.
- Github issues - Interested parties can comment on open issues or contribute their own tickets related to Attribution here.
- Bi-weekly Attribution community meeting - Meeting takes place every other Thursday at 1p CT. See rolling meeting notes.
- See the policy developed for the National COVID Cohort Collaborative (N3C) project: Attribution and Publication Principles for N3C published on Zenodo - doi 10.5281/zenodo.3992394.

8.6 Takeaway List

For individuals:

1. Identify contributors and track any short- or long-term roles on the project.
2. Establish contributors' roles in advance.
3. With respect to authorship, be transparent and clear about expectations and credit.
4. Use persistent identifiers!
5. Collect information about contributors as the project is launched and new people join a project.

For groups:

1. Incorporate CRediT/Contributor Role Ontology (CRO)/Contribution Attribution Model (CAM) into local workflows.
2. Provide opportunities for faculty and scholars to communicate their contributor roles on non-paper outputs.
3. Offer clear guidance on promotion and in faculty tenure documentation on how to incorporate contributor roles into their packet.
4. Likewise, publishers and funders should provide clear guidance as to how author contributions should be described for maximum effectiveness.
5. Provide feedback to the CRediT/CRO/CAM to request any missing roles that are not represented in the ontology/data model.

8.7 Deep Dives

For Individuals:

1. **Identify contributors and track any short- or long-term roles on the project.** This can be tracked on a project website or a collaborative online document (like a Google doc or a GitHub repository). Project websites offer a way to provide acknowledgment to project collaborators, especially for those who may not be an author on a resulting paper.
2. **Establish contributor's roles in advance.** Define clear expectations of roles and outputs for the project.
3. **With respect to authorship, be transparent and clear about expectations and author order.** The 'Guidelines on Authorship' from the University of Cambridge state "authorship criteria should be agreed by all investigators at an early stage of the research." [ref] Project leadership should provide friendly low-pressure opportunities for group and confidential discussions.
4. **Use persistent identifiers!** Please refer to the Best Practices Playbook chapter on PID (link) for a more comprehensive discussion on the topic, as well as quick takeaways including ORCID (www.orcid.org) for people and the preferred PID for a given topical domain or research community.
5. **Collect information about contributors as the project is launched and new people join a project.** This makes it easier to follow good practices and credit contributions in advance of paper submission or deposit of digital files into a repository. Suggested attributes to collect include: affiliation with the Reserch Organization Registry (ROR), preferred name, ORCID ID, grant numbers.

For Groups:

1. **Incorporate CRediT/Contributor Role Ontology (CRO) (https://data2health.github.io/contributor-role-ontology/) Contribution Attribution Model (CAM) (https://contributor-attribution-model.readthedocs.io/en/latest/) into local workflows.** This can be done collaboratively with stakeholders (e.g., thought leaders, system owners, community partners) and should offer opportunities for education about contributor roles, the importance of attribution, as well as provide an opportunity for feedback from stakeholders.
2. **Provide opportunities for faculty and scholars to communicate their contributor roles on non-paper outputs and provide context with their contributor roles on these items.** These include study materials, training and educational content, surveys, etc and the specific roles they played in generating these research outputs. See the Contribution Attribution Model (CAM) (https://contributor-attribution-model.readthedocs.io/en/latest/) for more detail.
3. **Offer clear guidance in promotion and tenure documentation to faculty on how to incorporate contributor roles into their packet.** If non-traditional scholarly outputs are recognized, these should be mentioned. This should be accompanied by real-life examples.
4. **Likewise, publishers and funders should provide clear guidance as to how author contributions should be described for maximum effectiveness.** Many publishers currently use the CRediT taxonomy for describing author roles. We recommend extending this to include the roles in the Contributor Role Ontology.
5. **After using attribution tools and best practices described here, scholars and organizational representatives should provide feedback to the CRediT/CRO/CAM to request any missing roles that are not represented in the ontology/data model.** This can be done via our GitHub issue tracker here: https://github.com/data2health/contributor-role-ontology/issues.

8.8 Contributors to this guidebook chapter

Contributor roles per CRediT or CRO

- Nicole Vasilevsky, Oregon Health Science University, 0000-0001-5208-3432, CREDIT_00000013 writing original draft role
- Lisa O'Keefe, Northwestern University, 0000-0003-1211-7583, CRO:0000065 project management role
- Kristi Holmes, Northwestern University, 0000-0001-8420-5254, CREDIT_00000013 writing original draft role

8.9 Acknowledgments

- CRediT - Contributor Roles Taxonomy. CASRAI.

8.10 Relevant Resources

Papers: Ilik V, Conlon M, Triggs G, White M, Javed M, Brush M, Gutzman K Essaid S, Friedman P, Porter S, Szomszor M, Haendel MA, Eichmann D and Holmes KL (2018) OpenVIVO: Transparency in Scholarship. Front. Res. Metr. Anal. 2:12. doi: 10.3389/frma.2017.00012

Pierce HH, Dev A, Statham E, Bierer BE. Credit data generators for data reuse. Nature. 2019 Jun;570(7759):30-32. doi: 10.1038/d41586-019-01715-4. PubMed PMID: 31164773. Available at Nature

Presentations: Credit Statement for the Force2019 Architecting Attribution Poster. DigitalHub. Galter Health Sciences Library & Learning Center, 2019. [doi:10.18131/g3-njgs-g416]((https://digitalhub.northwestern.edu/files/91c26739-87b5-407d-a1be-0f3a609a607a)

How to Enhance Attribution to Make More Meaningful Connections for Everyone to Their Roles, Work, & Impact. DigitalHub. Galter Health Sciences Library & Learning Center, 2019. [doi:10.18131/g3-y9vt-7376]((https://digitalhub.northwestern.edu/files/d08374e9-0411-4450-a0d1-4979c69ed3e7)

People + Technology + Data + Credit: Developing a Sustainable Community-driven Approach to Attribution. DigitalHub. Galter Health Sciences Library & Learning Center, 2019. doi:10.18131/g3-vs3n-ry93

Team Scientists: How Do We Enable Everyone to Get Credit for Their Work?. DigitalHub. Galter Health Sciences Library & Learning Center, 2019. doi:10.18131/g3-9q7s-5y55

Giving Credit Where It Is Due: How to Make More Meaningful Connections Between People, Their Roles, Their Work and Impacts. DigitalHub. Galter Health Sciences Library & Learning Center, 2018. doi:10.18131/g3-kqrj-z731

The Informatics of Attribution: a story of culture + technology in "New Ways of Counting Researcher Contribution" panel at the Society for Scholarly Publishing meeting. Washington, DC. 25 Sept 2018. Panel members: Casey Greene, Integrative Genomics Lab, University of Pennsylvania; Dina Paltoo, National Institutes of Health; Kristi Holmes, Northwestern University; and Vincent Lariviere, PhD, University of Montreal. Available at https://digitalhub.northwestern.edu/files/3db5f470-c519-46a9-9df3-357cf5d69a28

Understanding & Enabling Impact in the (with the) Community. Transforming Research Conference, Brown University, Providence, RI. 4 October 2018. Available at https://digitalhub.northwestern.edu/files/c6f785cf-f992-44ea-9905-ab4316181d91

Giving credit where it is due: how to make more meaningful connections between people, their roles, their work and impacts. FORCE2018, Montreal, Canada. 11 October 2018. Available here

Making it count: A computational approach to attribution. IEEE eScience Workshop on Research Objects (RO2018), Amsterdam, Netherlands. 29 October 2018. Available here and here

8.11 Funding:

This work was supported by the National Institutes of Health's National Center for Advancing Translational Sciences CTSA Program Center for Data to Health (Grant U24TR002306).

CHAPTER 9

Indices and tables

- genindex
- modindex
- search