Informatics playbook Documentation

CD2H Data Working Group

Nov 08, 2022

CONTENTS:

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

CHAPTER

ONE

CHAPTER 1: LICENSING FOR RESEARCH DATA

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

TWO

CHAPTER 2: BEST PRACTICES FOR USING IDENTIFIERS

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

THREE

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

FOUR

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

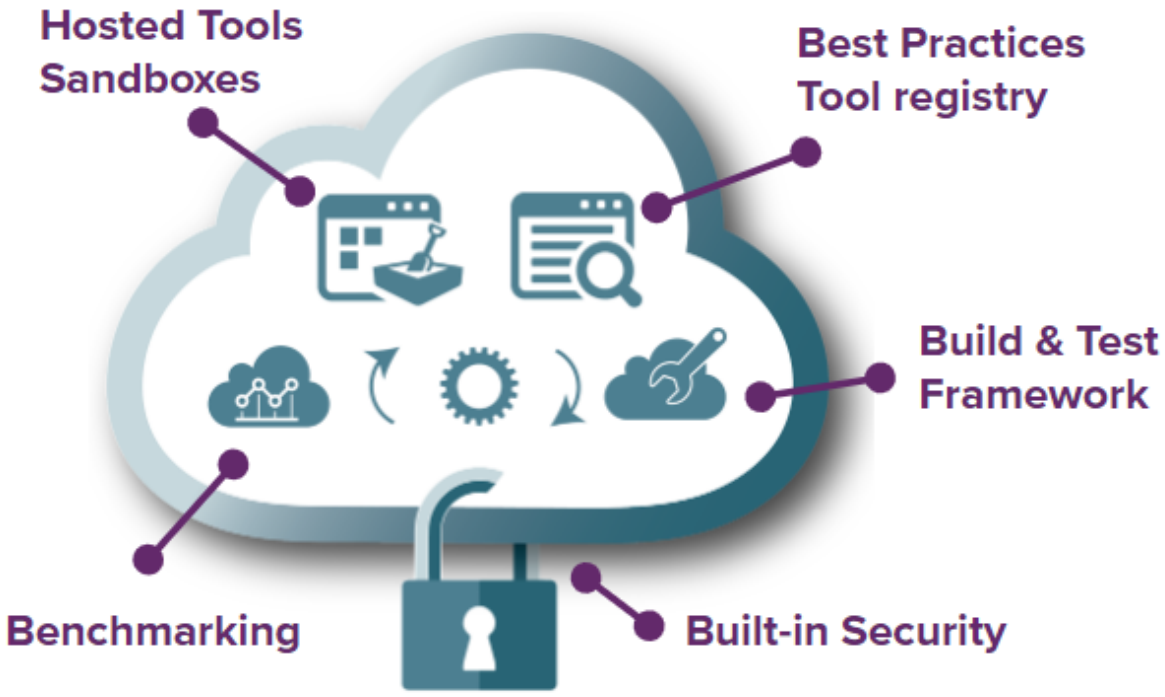
Thanks to Justin Ramsdill for the Agile introduction (also linked from the MTIP tutorial site).

CHAPTER

FIVE

CHAPTER 5: SOFTWARE AND CLOUD ARCHITECTURE

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 CD2H
ii. **Tool registry** to assist in the sharing and quality assurance of shared␣
↪software components developed by CTSA hubs     **\&lt;LINK TO SLIDES RE: TOOL␣
↪REGISTRY PROJECT\&gt;**
iii. **Build and test framework** for collaborative software development projects
iv. **Sandboxes** to provide spaces for informatics-focused workgroups seeking␣
↪solutions to shared data analytic and management challenges
v. **Benchmarking** of algorithms and predictive models using Challenge framework
```

1. **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 Form B. **CD2H cloud resource request "intake" form** (process for requesting access to CD2H provisioned cloud infrastructure) i. Cloud resource request intake form ii. 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 Labs E. **Benchmarking projects** leverage Challenge framework: i. Metadata Challenge (sharing of cancer-focused datasets) ii. EHR Challenge (mortality prediction)
2. **Takeaway list:** A. Create a common cloud computing architecture that can enable the rapid deployment and sharing of reusable software components by CTSA hubs B. 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 models D. Pilot the "cloudification" of software artifacts that can be shared across CTSA hubs to address common and recurring information needs.
3. **Deep dive into takeaways:** A. CD2H-NCATS Cloud Deployment Checklist B. CD2H-NCATS Cloud De-

ployment Process Workflow C. CD2H-NCATS Architecture Design Proposal D. CD2H-NCATS Architecture Request for Feedback Form E. CD2H-NCATS Federated Authentication (UNA) Overview F. 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

4. **Acknowledgements**
5. <LIST CLOUD CORE PARTICIPANTS>

5.1 Cloud Collaboration software

5.2 Cloud architecture

5.3 Software best practices

CHAPTER

SIX

CHAPTER 6: UNDERSTANDING DATA HARMONIZATION

6.1 About this document

6.1.1 Version: 1.0

6.1.2 Date: 4.23.20

Drafted and Edited by:

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

6.1.3 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.

6.1.4 Role of the CD2H Sustainability & Change Management Task Team (SCM)

CTSA informatics researchers strive to harmonize or combine their clinical data so that it can be available for users to query and view it in a unified format. The CD2H Sustainability and Change Management Task Team (*SCM*) of the Clinical *Data Harmonization* Working Group aims to aggregate, develop, curate, and disseminate content and tools that help encourage and guide organizations' understanding of and investments in *Data Harmonization* efforts. Many institutions are engaged in such efforts, but decision-makers do not always have a shared language for discussing these issues with ease or a full understanding of their facets. We aim to remedy that here.

6.1.5 Why is this work 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.

6.1.6 About the authors and contributors of this guide

The National Center for Data to Health (CD2H) tasked the *SCM* of the Clinical *Data Harmonization* Working Group with creating a maturity model for clinical *Data Harmonization*, as well as a roadmap and resources to help organizations plan and execute their harmonization efforts. The *SCM* aims to create educational events and provide information about the goals and benefits of *Data Harmonization*.

6.2 1. Why are we talking about Data Harmonization?

6.2.1 What is Data Harmonization?

Data Harmonization is the process of integrating disparate data of varying types, sources, and formats. *Data Harmonization* for research use refers to the ability of an organization to connect with external data-driven clinical and translational research projects involving patient data across multiple institutions (Clinical Data Research Networks or *CDRN*s, such as *PCORnet*, *ACT*, *OHDSI*, *TriNetX*). Coordinating this process involves not only the development and implementation of technology but also implementation of procedures and governance policies, and harnessing relevant expertise.

It is worth noting that the terms *harmonization* and *standardization* are both used to attain the same goal of data homogeneity. However, while standardization is about conformity, harmonization implies consistency. According to the Cambridge Biomedical Research Centre (UK), "standardization refers to the implementation of uniform processes for prospective collection, storage and transformation of data.[. . .] Harmonization is a more flexible approach that is more realistic than standardisation in a collaborative context. Harmonization refers to the practices which improve the comparability of variables from separate studies, permitting the pooling of data collected in different ways, and reducing study heterogeneity".*1* We can see the standardization approaches in the development of *Common Data Models (CDMs)* and harmonization becomes essential for retrospective collaborative research.

*CTSA*s have been building, participating in, and using *CDRN*s for several years with a variety of experiences and outcomes. The Clinical *Data Harmonization* Working Group believes these experiences contain lessons learned and intends to develop best practices for *Data Harmonization* from them. These best practices can guide organizations on emerging opportunities, such as new technologies and opportunities to join new networks. Helping develop, curate, and communicate/encourage action on this guidance is the purpose of the *SCM*.

6.2.2 Value Proposition of Data Harmonization

The quality, depth, and breadth of data and samples collected by healthcare organizations has provided increasing opportunities to advance knowledge and population health. Unfortunately, many healthcare institutions exist as silos with no ability to integrate patient data externally. Millions of inpatient and outpatient medical records containing structured data on clinical procedures, medications, diagnoses, demographics, lab results, and genomics, as well as unstructured data documented as free text in progress notes, cannot be used to their full potential. For example, rare diseases research or outcomes research, often require larger datasets than what is available at a single health care organization. Massive datasets offer more comprehensive analysis and the ability to probe specific research questions with more confidence in the results. To be able to make sense of vast information, *Data Harmonization* is needed to ask and answer important questions, and to provide evidence for policy changes in areas of standard of care, care delivery, coverage structure, and emerging population health trends.

In the past decades, the field has significantly moved from simple healthcare data digitization*2* to proactive and focused acquisition, aggregation, and standardization through the use of *Common Data Models* (*CDM*s). Currently, informatics teams frequently use harmonization and aggregation to improve the interoperability and compatibility of data among independent organizations that are willing to share and collaborate.*3* *Data Harmonization* is a stepping stone to conducting reproducible, collaborative research and is crucial to promoting efficient and timely communication between different stakeholders in the healthcare data domain. If not appropriately addressed, the lack of harmonization becomes a bottleneck, both for research and for sustainable operations. Impactful global health trends may be missed due to the variety of barriers, most of which lay within technical domains, as well as harmonization and standardization domains. A significant percentage of healthcare spending is wasted as a consequence of data silos*4*; therefore, the economic impact of data interoperability should be a significant factor in proposals urging for changes towards *Data Harmonization*. A significant effort has been made in *data standardization* and harmonization of research datasets through the work of Clinical Data Interchange Standards Consortium (CDISC).*5* It is time for healthcare data used for research to undergo a similar transformation.

SUMMARY: Data Harmonization *has a direct impact on quality, reproducibility, and confidence of research that uses data contained in the EHR system. Effective data interoperability leads to improved efficiency of methods and processes, improved time from research project design, through trend identification to policy implementation, and results in cost savings due to increased efficiency and a decreased need for data wrangling. Emerging local, national, and global networks, such as* ACT**, PCORnet**, OHDSI**, TriNetX *and others that rely on *Common Data Models *and harmonized data sources, bring about opportunities for global-scale research and policy implementation. Analysis of massive harmonized datasets can provide better insight into population health and reduce local, national, and global healthcare spending waste.*

6.2.3 Risks of not moving forward

What do research enterprises risk if they do not participate in *Data Harmonization*? A lack of harmonization restricts researchers to analyzing different data sources separately, which is counterintuitive to the reality of present-day collaborative research. Project-specific *Data Harmonization* is an exhaustive, time-consuming process*6*. Reports indicated that research project harmonization can require up to 6 months to complete, and that about 30% of the data deemed impossible to harmonize and was excluded from the final dataset. Time spent on wrangling data and custom coding takes resources away from finding insights that drive discoveries forward and slows the process of achieving public benefit from them. This exhaustive process negatively impacts the time to and cost of discovery and, therefore, the cost of care. The cost increase eventually creates a barrier for access to care by disadvantaged segments of the population and smaller, less developed countries.

Pending a comprehensive literature review, it is evident that the risks of not moving forward with *Data Harmonization* efforts include, but are not limited to: (1) the extensive use of resources (both human and financial) on data wrangling, rather than on research and implementation of findings into healthcare; (2) the delayed ability or inability to improve human health based upon the knowledge that is gleaned from larger, harmonized disparate datasets; and (3) the reduction of an institution's competitiveness for grant funding and top faculty recruitment, which could cause decreased pharmaceutical engagement for high profile studies.

SUMMARY: The risks of not moving forward with Data Harmonization *lead to protracted research efforts, delays in translating research to improve population health, and creates barriers for access to care for the disadvantaged.*

6.2.4 Assessing the impact of Data Harmonization

To fully understand the impact of the *Data Harmonization* efforts on areas of research operations, one needs to define, implement, and consistently collect metrics to assess the return on investment (ROI) of those initiatives. If a research enterprise chooses to embark on *Data Harmonization* activities, how do we measure the ROI and impact? What are the potential metrics of success? How is success characterized?

As we will demonstrate below, *Data Harmonization* is a part of the measure of organizational maturity, and achieving a high score on the maturity index means the organization is implementing strategies identified as impactful for the overall success. Regularly assessing organizations on the maturity index for the *Data Harmonization* domain can provide insights into the success of the effort. Additionally, informatics teams can implement processes and research-impact related measures alongside the maturity index. These key performance indicators (KPIs) could include, but are not limited to, the following examples; *(1) the decrease in time required for* Data Harmonization *steps; (2) the percent increase of data available for research after the harmonization process (i.e. decrease in discarded data because of the inability to harmonize); (3) the increased number of multi-site research and collaborations; (4) the increased number of publications, findings, and policy implementations; (5) intellectual property filings that can be attributed to organizations choosing to invest into* Data Harmonization *efforts either directly or toward initiatives leading up to it.*

6.3 2. Is your organization ready for Data Harmonization?

Maturity refers to the organizational capacity to deliver a service while considering multiple factors including culture, policy, and organization. The *Data Harmonization* maturity index is a questionnaire that assesses an organization's capacity to connect with external data-driven clinical and translational research projects (often referred to as Clinical Data Research Networks or *CDRN*s) involving patient data.

6.3.1 Fill out the Data Harmonization Maturity Index for Assessment

Click here (http://j.mp/2GvfPum) to go to a draft version of the *Data Harmonization* Index. This index is under development; however, even in draft form, it can provide insights into an organization's maturity regarding mission, governance, sustainability, workforce, infrastructure, relationship with the clinical enterprise, data practices, and external relationships. It also provides topics for an institution to discuss regarding its capacity to participate in *Data Harmonization* efforts.

After reviewing the results, read the section titled "Harmonized Data Maturation" to learn more about aspects of the maturity assessment.

6.4 3. Harmonized Data Maturation

7 Data Harmonization is a journey, not a destination. To realize its values and benefits, organizations must make it an integral strategic component of daily practices and workflows. Here we discuss 8 maturity categories that are essential for harmonized data maturation and sustainability. As organizations increase their capabilities in each area, they increase their organizational maturity.

6.4.1 Data Harmonization Mission

Maturity in this category means:

- Organization has defined a reason to participate in a Clinical Data Research Network (CDRN)- Organization has defined what CDRNs they want to participate in
- Organization has an implementation plan for participating in a CDRN- Organization has a strategic plan for participating in a CDRN **Defining Participation Reason.** Understanding the current state of the biomedical academic enterprise provides insight into the "why" of *Data Harmonization* and the challenge of sustaining harmonization efforts. There are extensive databases of ever-expanding clinical data due to the national uptake of electronic health records in hospitals and clinics of all sizes. Most institutions have also created data warehouses, sometimes several, for specific purposes (e.g., operations or research), out of which researchers can export EHR and other data for long-term storage and access them for purposes other than clinical care.

Computer science algorithms have been created to process these data for many types of research. Some methods collectively, although obtusely, referred to as "AI" include machine learning and data mining techniques, and are being applied to further develop vast data sets. Many researchers are working to demonstrate the value that these computational methods can provide to health care. Areas of interest include "development of drug development pipelines; phenome-wide association studies when linked with genotyped biospecimens; characterization of treatment pathways for large numbers of patients, multi-site trial recruitment, automated data collection, and delivery of real-word evidence at the point of care to actualize the learning healthcare system, e.g., developing predictive models for clinical decision support put into place within the EHR for a variety of conditions."*8* Pooling large amounts of clinical data may make major research strides possible. Technologically, we can handle the storage and processing power needed to do so, although there are costs to all institutions to ensure this.

An organization must support clinical *Data Harmonization* via implementation and ongoing maintenance of crosswalks of mapped data, which requires informatics expertise and guidance. These mappings or crosswalks connect different *Common Data Models* (*CDM*) to ensure that each data element from one *CDM* matches with the correct data element from another *CDM*. These crosswalks provide interoperability and support reproducibility.

Defining CDRN Participation. Researchers at an organization must be able to easily determine which *CDM* and *CDRN* their organization participates in, which data this connection will afford them to access, and what additional benefits they can receive because of the *CDRN* participation. To support this goal, organizations must be able to state which research networks they will and will not participate in, the particular *Common Data Models* they do not support, and what data their researchers will not have access to as a result. They also must understand the choices their collaborators and competitors are making regarding *CDRN* and *CDM* and the potential impact of those choices. Finally, they need to understand how their choices may impact the organization's credibility, their standing in the research world, the opportunities for their researchers, and their ability to recruit competitively.

Creating a CDRN Implementation Plan. Why is *Data Harmonization* necessary for interoperability in the first place, especially given that so many institutions have purchased their EHR from the same vendor?

There are myriad ways that the set-up of these systems can affect which data are collected, at what level of granularity, how they are labeled, and the concepts for which any given piece of data is meant to represent. In addition, when data are exported from their source system, e.g., an EHR, into a data warehouse, decisions are made that can affect the data in many ways (such as the impact of ETL based on decisions and the rationale behind them.

Consortia of institutions have been formed to work together to create mappings for harmonizing their data and sharing it in Clinical Data Research Networks (*CDRN*). The most well known include the Accrual to Clinical Trials (*ACT*) network; the All of Us research program; *TriNetX*, a private company-driven network to support clinical trials; and *PCORnet*®, National Patient-Centered Clinical Research Network, a group of networks including 9 clinical research networks, 2 health plan research networks, several coordinating centers, and its main funder, PCORI, the Patient-Centered Outcomes Research Institute.

Each participating institution maps their data to a *Common Data Model (CDM)* developed or used by a *CDRN*. In the past, institutions had to build their models from scratch. Sometimes they still try to do this and end up with the models very similar to the existing *CDM*s. The creation of a new *CDM* is therefore rarely needed. A widely used *CDM*

optimized for retrospective observational studies is the Observational Medical Outcomes Partnership (*OMOP*) *CDM* https://www.ohdsi.org/data-standardization/the-common-data-model/, which has gained traction among data scientists who also conduct machine learning. The All of Us research program employs *OMOP* as its *CDM*. *PCORnet* has its own *CDM* https://pcornet.org/data-driven-common-model. **Integrating Participation into Strategic Planning.** Adequate funding is essential to *Data Harmonization* maturation, as are strategic efforts to ensure that when we exchange and pool data from our data warehouses, it matches with "like" data from other institutions. Depending on which national research networks an institution wants to participate in, *Data Harmonization* for those networks will have to be supported.

Because each *CDM* requires ongoing additions to concepts, data types, other changes, updates to the model itself, and to dependencies (e.g., searches, interfaces, documentation, training, etc.), their maintenance requires ongoing expertise and effort. Institutions must commit to the level of *CDM*s and *CDRN* participation they can sustain. An organization that is mature in its *Data Harmonization* efforts can handle more models and more complexity.

While few argue against the merits of *CDM*s, demonstrable differences in content and structure among them are becoming burdensome for many *CTSA* to manage. The emergence of *FHIR* as a clinical standard may mitigate these concerns through a model of low-cost, low-effort creation of *FHIR*-based repositories from APIs that will soon be native to all EHRs, per pending regulation arising from the 21st Century Cures Act. These *FHIR* repositories, by design, would embrace a canonical *data model* framed on the US Core *FHIR* Implementation Guide. Given a canonical data hub, distributing algorithmic transformations, created and maintained by consensus, would allow *CTSA* hubs to derive data marts and projections conforming to popular *CDM*s. Further, *federated queries* across these hubs may obviate the need for research-specific *CDM*s, though that remains speculative in the face of the current maturity level in *FHIR*. Data exchange harmonized with *FHIR* can facilitate streamlined integration of data sources for improved population health research.

6.4.2 Governance

Maturity in this category means:

- Organization has a high-level group that effectively makes decisions about data, including external data sharing agreements.
- Organization's IRB is experienced in reviewing protocols that use CDRNs.
- Organization has the legal processes in place to effectively enter into external data sharing agreements.
- Organization's processes enable decision making at appropriate levels.

An organization must have a governance infrastructure group that makes decisions about data, including *Data Harmonization* and research network participation. An organization must have a governance group to make decisions about data and participation, including external data sharing agreements, and participation in *CDRN*s, and *Common Data Models* (*CDM*s). The organization's Institutional Review Board (IRB) must be proficient in reviewing protocols that use *CDRN*s. The organization must have legal processes in place to effectively enter into external data sharing agreements. Organizational processes must enable decision-making at appropriate levels.

6.4.3 Sustainability

Maturity in this category means:

- Organization has made a commitment to support Data Harmonization.- Organization has identified internal funding sources to provide dedicated resources for Data Harmonization.- Organization has obtained external grants to fund projects that use harmonized data.
- Organization has an effective recharge model in place to help recover some of the costs of Data Harmonization for funded research projects.

An organization must define why it is participating in one or more Clinical Data Research Networks and associated *Common Data Models*. It must declare ongoing organizational and financial support to sustain the *Data Harmonization* necessary for participation. Organizations cannot sustain harmonization without alignment with their mission and ongoing declared and financial support. This is true for all infrastructure related to conducting research. EHR data provide a rich source of information, and various internal and external entities continue to request access in large numbers. However, the existence of data in the EHR does not mean it is automatically accessible or "shovel-ready" to use for secondary purposes. Making EHR data accessible and ready to use requires infrastructure and expertise. One of many data-related considerations in *Data Harmonization* is the need to prepare data for interoperable use across organizations via the implementation of *Common Data Models*, most often in Clinical Data Research Networks (*CDRN*s). A useful 2014 article, "Sustainability Considerations for Health Research and Analytic Data Infrastructures," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4371522/) discusses this topic and the reasons why leaders must address data infrastructure sustainability.

Institutions must allocate funds for *Data Harmonization* efforts and incorporate them into capital and operating budgets. One of the ongoing entrenched challenges is the establishment and funding of a health information technology (HIT) infrastructure to support research. Of specific importance is establishing who is responsible for funding the *Data Harmonization* efforts. Deans of institutions favor NIH funding for research related activities; however, NIH does not focus on sustaining IT or HIT infrastructure or determining and supporting data standards. Stanford University's John Ionnidis, MD, DSc, and others have spoken eloquently about funding challenges and misalignments in the scientific enterprise, and academic medical centers' over-reliance on soft money.*9*,*10*,*11* Enterprise IT departments often set capital budgets and strategic plans that encompass hardware, software, and tangible deliverables. However, these often do not cover the scientific expertise and personnel necessary to work on data modeling and *data mapping*. Although *CTSA* and PCORI awards can be used for specific projects and purposes regarding *Data Harmonization*, they are not intended to replace allocated funds from within the institution for sustaining infrastructure and the human expertise that must accompany it. If institutions want to be competitive in research, and compete regarding the external use of their de-identified data for commercial purposes, then committing to *Data Harmonization* support is necessary. Benchmarking with other institutions is key for developing appropriate ongoing financial investment and support for *Data Harmonization*.

6.4.4 Workforce

Maturity in the category means:

- Organization has a team dedicated to curating and maintaining data in a manner that effectively enables sharing data with CDRNs.
- Organization has data architects to work on Data Harmonization efforts.
- Organization has clinical informaticists to work on Data Harmonization efforts.
- Organization has a business analyst to work on Data Harmonization efforts.
- Organization has project managers available to work on Data Harmonization efforts. An organization must determine how many staff and what type of expertise are necessary to implement and maintain *CDM*s for the organization's declared *CDRN* participation and interoperability goals. Although partners in *CDRN*s use automated processes to extract, transform, map, and standardize data to the selected *CDM*s, the process requires ongoing human expertise, oversight, curation, and quality assurance testing to ensure that the data are correct, complete, and the processes are working. This doesn't happen by itself, nor will it ever, and will require people with expertise to oversee and conduct these processes.

The organization must determine which staff will maintain these data, how many are required to perform the work, and what their work entails. Examples of tasks associated with maintaining data include participation in ongoing consortial meetings, following up on updates, extending mapping to cover additional data, and performing required quality checks.

Benchmarking with other organizations is critical in determining adequate and optimal staff who possess the level of expertise that highly successful *Data Harmonization* teams include, e.g., informaticists. Towards this effort, organi-

zations should consult the in-progress maturity model for *Data Harmonization*. Additionally, they should track the efforts of ongoing research projects that address workforce development issues (e.g., 2019-2020 deliverables stemming from the *CTSA* Informatics EC and efforts of its Enterprise Data Warehouse Working Group). These activities will shed light on identifying right-sized and right-type staffing.

A final, but not insignificant, workforce consideration is the need for experts to provide strong support and training to aid end-users and ensure responsible use of these harmonized data. Whether those people come from the *Data Harmonization* team or not, they must train and support end-users of data by ensuring they share the right type and amount of details regarding *CDM*s and *CDRN*. End-user training should also include discussions of participation and access issues, and ultimately, increase end-user data literacy.

6.4.5 Infrastructure

Maturity in this category means:

- Organization has sufficient hardware resources (including networking, storage and server) to enable Data Harmonization efforts.
- Organization has sufficient software resources available to enable Data Harmonization efforts.
- Organization has an IT infrastructure that supports Data Harmonization efforts as part of the organization's enterprise IT infrastructure. To fully support *Data Harmonization* with enterprise IT infrastructure, an organization's IT Department needs to act explicitly towards that end. Enterprise IT must take into account the actual hardware, networking, storage, server, and other software requirements of the institution's *Data Harmonization* work and personnel, and ensure that these needs are planned for and met in ongoing planning and budgeting processes.

6.4.6 Relationship with the clinical enterprise

Maturity in this category means:

- Organization conducts clinical enterprise values research.
- Organization has an effective clinical IT services operation.
- Organization has a clinical IT reporting team that is aware of Data Harmonization efforts.
- Organization has a clinical operation that is aware of the secondary use of clinical data to drive medical research.

This category is related to the research value of the clinical enterprise, such that there is a relationship between the clinical enterprise and *Data Harmonization* efforts. The clinical data operation will be aware of the secondary use of clinical data to drive medical research. Most organizations have established this concept, but depth of understanding of data issues and *Data Harmonization* varies widely, and "data literacy"—what that means and how to develop it—among clinical data users and researchers, will need ongoing attention. In a mature organization pursuing *Data Harmonization*, the clinical IT EHR reporting team is aware of *Data Harmonization* efforts and can speak cogently about them.

6.4.7 Data practices

Maturity in this category means:

- Organization's Data Harmonization activities are consistent with enterprise data practices.
- Organization's Data Harmonization teams have established an aggregated database that feeds Data Harmonization efforts.
- Organization follows good service management methods for Data Harmonization efforts.
- Organization tracks data provenance for data shared with CDRNs.
- Organization practices FAIR (Findable, Accessible, Interoperable, Reusable) Principles in managing data shared with CDRNs.

For sustainability, best practices for data must be followed and include *Data Harmonization* activities that are consistent with enterprise data practices. The organization must follow good service management methods for *Data Harmonization* efforts, as well as FAIR (Findable, Accessible, Interoperable, Reusable) Principles for managing data shared with *CDRN*s. The organization will track data provenance for data shared with *CDRN*s. The organization's *Data Harmonization* team will have established an aggregated database that feeds *Data Harmonization* efforts.

6.4.8 External relationships and outreach

Maturity in this category means:

- Organization has strategies for establishing external collaborations
- Organization is experienced in working with multi-institutional research collaborations
- Organization has participated in at least one clinical data research network

The purpose of *Data Harmonization* is the interoperability and exchange of data, particularly for research; therefore, developing, participating in, and maintaining external relationships and outreach for data exchange is an inherent sustainability goal. Organizations must develop strategies for establishing external collaborations, and demonstrate experience with working on multi-institutional research collaborations, including at least one clinical data research network.

6.5 4. Data Harmonization Technologies

All *CTSA* hubs understand the basic advantage of a *federated query* and its underlying requirement for a *common data model.* Most hubs participate in multiple efforts for standardizing data into multiple *Common Data Models*, which become increasingly less scalable. Meanwhile, the clinical community has been grappling with the interoperability problem, and are seeking a common framework for clinical data rendering and exchange. There has been dramatic progress in clinical data standards over the past 5 years, which provides an opportunity to leverage emergent and dominant clinical data standards.

6.5.1 Federated Query

A *federated query* provides the ability for one coordinating group to publish an executable algorithm, typically case and control cohorts with comparison analytics that can be independently executed by consortium members. The *CTSA* hubs function in this role. The *federated query* approach ensures that no patient-level data leaves an organization, thereby preserving confidentiality and obviating disclosure. Other more sophisticated models specify a matrix of tabulations, covariates, or features; a benchmark in the sequence to calculate parameter estimate and variance; and a merger of the matrices for more robust meta-analyses. Since the system only returns cell aggregate sub-totals in the matrix, patient-level data is not exposed.

6.5.2 Common data models

Widespread adoption of electronic health records (EHR) and emphasis on the reuse of clinical data through integration of clinical care, claims, environmental, and other data requires a robust approach to data modeling to satisfy the complexity and provide maximum usability for effective research. As summarized by Khan et al in the analysis of existing *data models* "Data modeling is the process of determining which data elements will be stored and how they will be stored, including their relationships and constraints. The structure and definitions of a data model define what data can be stored, how values should be interpreted, and how easily data can be queried."*12* Significant efforts that have been made by the research community to address the issue of standardization and effective data modeling resulted in a few prominent, widely accepted *Common Data Models* (*CDM*)—*PCORnet*, *OMOP*, *i2b2*, which is well described in a work by Jeffrey G. Khan and colleagues about data model harmonization for the "All Of Us" Research Program.*13* The section from the publication is shown below to help the audience clarify the landscape of the *Common Data Models*:

"PCORnet common data model (PCORnet *CDM*) The *PCORnet*t *Common Data Model* is supported by all networks in the Patient Centered Outcomes Research Institute, and thus has a wide base of existing support. Over 80 institutions have already transformed their data into this model. It was derived from the Mini-Sentinel *data model*, which has increasing uptake in claims data analysis.

PCORnet *CDM* (v3.1) is a traditional relational database design, in which each of fifteen tables corresponds to a clinical domain (e.g., diagnoses, labs, medications, etc.). The tables have many columns including both the table key (patient identifier, encounter identifier, etc.) and additional details (e.g., medication frequency). New releases of the *data model* have added new clinical elements or format–for example, new domains (e.g., lab values) and changes in data representation (e.g., smoking status).

Informatics for integrating biology in the bedside (*i2b2*) *i2b2* was first developed over a decade ago through a National Institutes of Health (NIH) grant and continues to grow in popularity. It is currently used at over 200 sites worldwide, and it is used in several large-scale networks, including the NCATS' national Accrual to Clinical Trials (*ACT*) network.*14*,*15*. *i2b2* uses a star-schema format, pioneered by General Mills in the 1970s and widely used in retail data warehouses.*16* The *i2b2* star-schema uses one large "fact" table containing individual observations. This is a narrow table with many rows per patient encounter. *Ontology* tables (hierarchical arrangements of concepts) provide a window into the data; these are often developed by local implementers. Consequently, the *data model* is only modified when core features are added to the platform. Observational Medical Outcomes Partnership (*OMOP*)

OMOP was developed to be a shared analytics model from the beginning, and it has been adopted by the Observational Health Data Sciences and Informatics (*OHDSI*, pronounced "Odyssey") Consortium, a diverse collaborative dedicated to research and quality improvement.*17* The *OMOP CDM* is increasingly utilized, presently at 90 sites worldwide, thanks to *OHDSI*s large community and many analytical tools. *OMOP* is a hybrid model that provides domain tables in the vein of *PCORnet*, as well as a "fact" table containing individual atomic observations similar to *i2b2*. The *OMOP* schema is significantly more complicated than *PCORnet*, and some domain tables are derived values for specific analytical purposes (e.g., drug_era and visit_cost). Unlike *PCORnet* (but similar to*i2b2*'s *ontology* system), *OMOP* provides metadata tables providing information on *terminology* and concept relationships."*18*

While each *CDM* has creatively executed solutions to support research, each also has its limitations. Organizations planning *Data Harmonization* efforts, implementing data warehouses, and choosing *CDM*s ultimately need to make a

decision based on the majority of the research needs, which has become increasingly difficult. More and more often, sites realize that in order to participate in multi-site research, they need to support all 3 data models to achieve data interoperability with a collaborator.

6.5.3 Interoperability: Clinical data rendering and exchange

Fast Healthcare Interoperability Resources (*FHIR*, pronounced "fire") is a vital standard related to interoperability created by the Health Level 7 (HL7) International health-care standards organization. This standard consists of data formats and elements ("resources") and an application programming interface (API) for exchanging electronic health records (EHRs.) *FHIR*'s greatest strength is that it comes natively out of EHRs through APIs that are regulatorially required.

FHIR is an open, collaborative community working harmoniously with healthcare, academia, governments, EHR vendors, payers, and patient groups to define a shared specification for health data exchange. It builds upon previous work, while retaining the flexibility to make specification easily usable and understandable. Many practical demonstrations have proven FHIR's effectiveness at interoperability in open connect-a-thons (an event to prove the efficacy of the standard).

Why does FHIR matter?

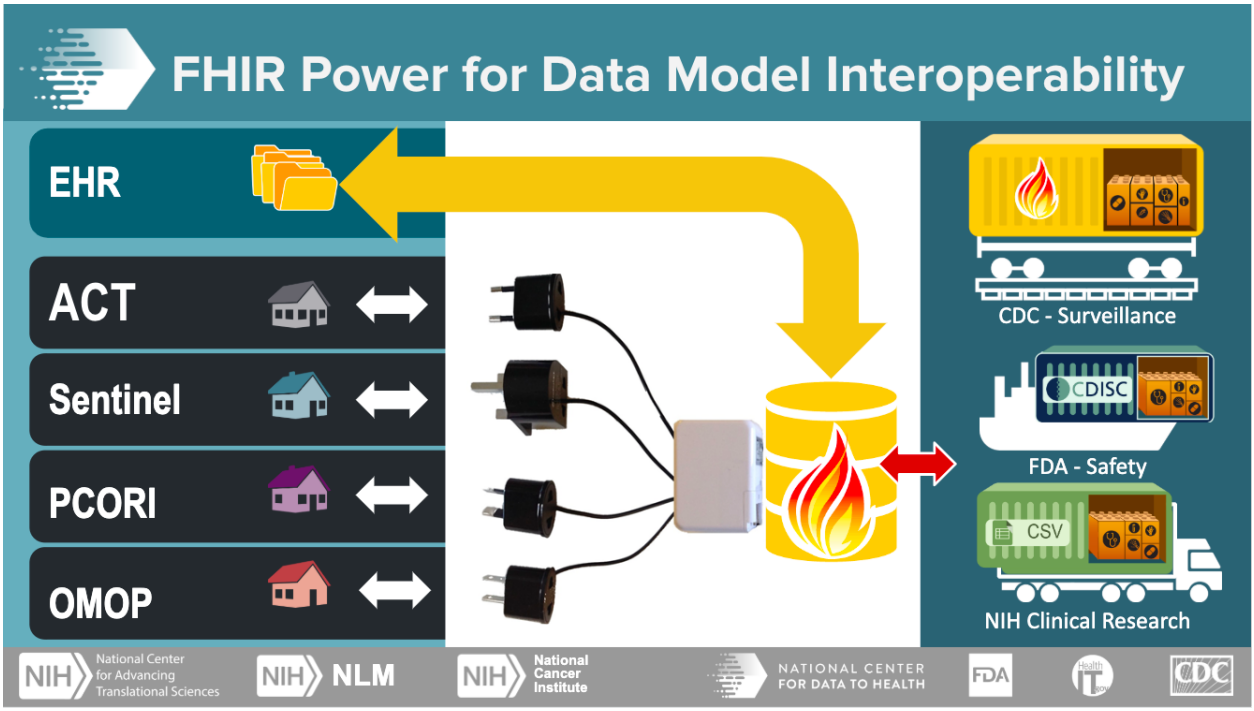
FHIR is unique among health information technology standards for enjoying nearly universally positive responses among health systems, academics, government(s), payers, pharma, and most EHR vendors. Many communities have created *FHIR* Accelerators to enrich the rapid progress and maturation around the standard. Most compellingly, *FHIR* is embraced and adopted by the system developers for whom *FHIR* was designed. This translates into a proliferation of *FHIR* throughout the clinical community.

The translational research community can benefit directly from the enormous investments these clinical communities using *FHIR* make. The workforce and investments dwarf any that could be marshaled by the translational research community. Rather than reinvent research-specific data standards, it makes sense to leverage the structure, specification, detail, tooling, and infrastructure coming forward from these clinical communities. Pending Office of the National Coordinator for Health Information Technology (ONC) regulations will require that all EHRs support *FHIR* API interfaces, profoundly simplifying data access and delivery in a *FHIR*-conformant manner. This effectively obviates any need for complex extraction and translation in the creation of *FHIR*-based repositories.

NIH is investing in *FHIR*, not only for EHR exchange but also for research from academia to industry. NIH encourages its researchers to explore the use of the Fast Healthcare Interoperability Resources (*FHIR*®) standard to capture, integrate, and exchange clinical data for research purposes and to enhance capabilities to share research data*19*,*20*. This use includes:

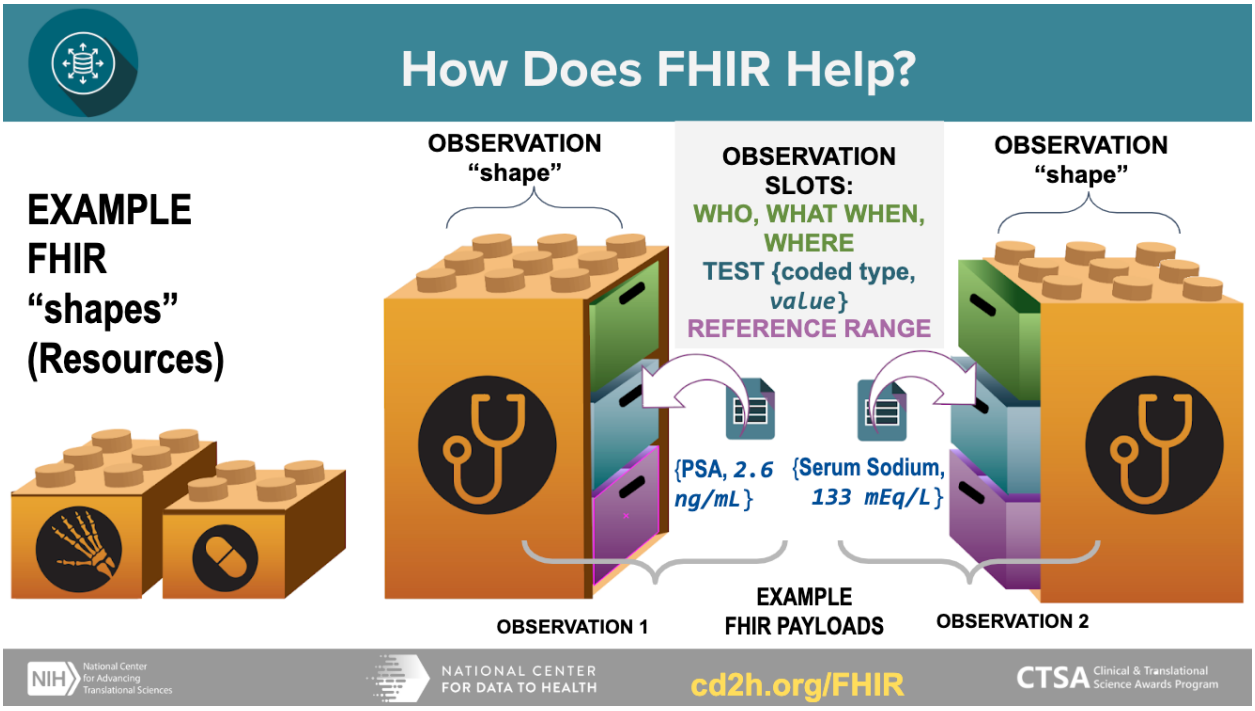
- Integration of patient- and population-level data from EHR systems
- Access to and management of electronic health information
- Development of clinical decision support systems
- Exchange of EHR data and health documentation
- Enhancement of privacy and security for electronic health information
- Inclusion of common data elements (CDEs) and clinical terminologies to improve interoperability of EHR data for research and clinical care
- Support of a common structure for sharing clinical research data
- Integration of EHR and patient-originated data with clinical research data
- Design and monitoring of clinical trial protocols
- Enhancement of patient recruitment, enrollment, and consent in clinical trials

NIH and related agencies, together with CD2H, have leadership positions in the newly established HL7 FHIR Accelerator for research called Vulcan. The goal of this effort is to ensure the needs and requirements of translational research, specifically those of the *CTSA* community, are accommodated and incorporated in *FHIR* development, evolution, and maturation.



How does FHIR work?

FHIR has modular data models that can carry a flexible payload (data), and reusable "containers" that can be assembled into working systems. This model enables the exchange of well-defined structured data in small discrete units.



6.5.4 Getting started with Data Harmonization technologies

There are several questions to consider when approaching tools and technologies for *Data Harmonization*:

- If an organization with a deployed EHR (Electronic Health Record) wanted to start from scratch using FHIR (Fast Healthcare Interoperability Resources), they should consider the following:
  - Any EHR (other than home grown) will already have a good amount of FHIR capability.
    - Pros of using FHIR: it is not dependent on specific EHR APIs and data structure, it has ability to leverage 3rd party tools and software, it can connect to other applications, and provides a gateway to research and general data sharing.
    - Cons of using FHIR: it masks the advanced features of a specific EHR—meaning there are restrictions on certain functionalities.
  - In general, an organization should interact with the EHR directly when necessary, but try to leverage FHIR as much as possible for "future proofing"
  - An organization should consult the official HL7 website https://www.hl7.org/fhir for instruction on installing FHIR services. - An organization will want to consider the following aspects to determine if using FHIR alone is sufficient justification for moving to the cloud:
    - FHIR and the cloud are orthogonal (independent). FHIR servers can be deployed locally behind firewalls on a custom machine, and they can also be deployed in a cloud environment.
    - Cloud based deployment packages (from specific vendors) help get stand-alone FHIR servers up and running more quickly, but there are non-cloud solutions as well.
  - An organization will also want to consider how FHIR services will affect the operation of the clinical database. Because analytic queries across a transactional database are needed for multiple patients, it is important to consider how this will affect performance.
    - Operation and performance are the function of how a specific EHR implements FHIR, and a secondary stand-alone FHIR server clinical data warehouse model can be used in cases where it is needed. Note:

All considerations should be discussed with the organization's EHR Administrator and Senior IT Directors. They can address current EHR set-up, workflow impact, and architectural considerations for implementing *FHIR*. EHR Administrator and Senior IT Directors may include:

* CIO - for overall information strategy
* Chief Medical Information Officer - for addressing issues where medicine meets technology
* Chief Security Information Officer - for security and privacy issues
* EMR Project Management Office (PMO) - for project planning and execution
* Sr. Director Infrastructure - for networks, servers, workstations, storage, and cloud
* Sr. Director Clinical Information Systems - for EMR plus other systems
* EMR Project Management Office (PMO) - for project planning, execution, usually reports to above
* Sr. Director Clinical Research - for research aspects of EMR usage
* Data Trust/Governance Chair - for use of data and standardization, which most organizations have

6.6 5. Appendix: Terminology

A B *C D* E *F* G H I J K L M N *O P* Q R S *T* U V W X Y Z

6.6.1 A

ACT Network

(See Accrual to Clinical Trials (ACT) Network)

Accrual to Clinical Trials (ACT) Network

The ACT Network is a collection of nearly 40 million patient records that are available to clinical researchers and study teams for the purposes of discovery, exploration, and validation of patient cohorts for investigator-initiated multi-site or single-site clinical trials. Provided by the University of Minnesota, this network can be used to help identify potential partner sites for multi-site studies. It is not necessary to have an established collaborator at another site to use the tool for cohort discovery.

Additional Resources: https://www.ctsi.umn.edu/consultations-and-services/multi-site-study-support/accrual-clinical-trials-act-network

6.6.2 C

CDM

*(See "***Common Data Model (CDM)***")*

CDRN

*(See "***Clinical Data Research Networks***")*

Clinical Data Research Networks (CDRN)

A Clinical Data Research Network (CDRN) is a group of institutions that agree to store their electronic health record data in the same format, thus harmonizing data elements to standardized definitions so that clinical data queries can be run across all member sites.

Additional Resources: https://tracs.unc.edu/docs/bmi/NC_TraCS_CDRN_Overview_20170426.pdf

Common Data Model (CDM)

Microsoft: The Common Data Model (CDM) comprises a standardized metadata system and data schemas that help applications and services interoperate and help you get more value from your data. The Common Data Model simplifies data integration and application-development scenarios, allowing independent development while maintaining shared meaning of data.

Examples of a common data model:

- **PCORnet***(**https://pcornet.org/data-driven-common-model/**)* This common data model allows EHR, registry and claims data to be structured the same way across many institutions, allowing researchers to answer questions across larger sets of data than their own institution, specifically to find cohorts for clinical trial participants. PCORI funded the development of PCORnet.
- **OMOP/OHDSI** * (https://www.ohdsi.org/ **, https://www.ohdsi.org/data-standardization/***)* The OMOP (Observational Medical Outcomes Partnership) common data model was designed for support of observational research by converting various types of observational data through the OMOP Common Data Model. OMOP expanded to become Observational Health Data Sciences and Informatics (OHDSI), the OMOP Common Data Model as well as a common representation (terminologies, vocabularies, coding schemes). A systematic analyses is performed on the CDM using a library of standard analytic routines that have been written based on the common format. The data from these different sources are converted into three types of evidence: clinical characterization, population-level effect estimation, and patient-level prediction.
- **I2b2***(**https://i2b2.cchmc.org/faq**)* I2b2's common data model is also used to find cohorts for clinical trials. It uses an ontology that maps data across institutions into facts and dimensions. A fact is the piece of information being queried, and the dimensions are groups of hierarchies and descriptors that describe the facts. *Additional Resources: https://docs.microsoft.com/en-us/common-data-model/faq#what-is-the-common-data-model*

Clinical and Translational Science Awards (CTSA) Program

Under NCATS' (National Center for Advancing Translational Sciences) leadership, the Clinical and Translational Science Awards (*CTSA*) Program supports a national network of medical research institutions—called hubs—that work together to improve the translational research process to get more treatments to more patients more quickly. The hubs collaborate locally and regionally to catalyze innovation in training, research tools and processes.

The *CTSA* Program is designed to develop innovative solutions that will improve the efficiency, quality and impact of the process for turning observations in the laboratory, clinic and community into interventions that improve the health of individuals and the public.

Additional Resources: https://ncats.nih.gov/ctsa **Controlled vocabulary**

A set of terms that are selected and defined based on the requirements set out by the user group, usually a set of vocabulary is chosen to promote consistency across data collection projects. These terms have a fixed and

unalterable meaning, and from which a selection is made when cataloging; abstracting and indexing; or searching books, journals and other documents and resources. The control is intended to avoid the scattering of related subjects under different headings. An example of a controlled vocabulary is the Getty Vocabularies http://www.getty.edu/research/tools/vocabularies/index.html.

Additional Resources: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C48697; *http://purl.bioontology.org/ontology/MESH/D018875*

CTSA

*(See "***Clinical and Translational Science Awards Program***")*

6.6.3 D

Data aggregation

Data aggregation is the compiling of information from databases with intent to prepare combined datasets for data processing. MeSH: Data aggregation - Process of searching, gathering, and presenting data in a summarized format.

Note: The difference between data aggregation and *Data Harmonization* is that data aggregation involves compiling information for processing into datasets, while *Data Harmonization* uses search and mapping technologies to offer views across disparate datasets.

Additional Resources: https://en.wikipedia.org/wiki/Data_aggregation#cite_note-1 OR https://meshb.nlm.nih.gov/record/ui?ui=D000078303

Data architect

A data architect primarily ensures that an organization follows a formal data standard and that its data assets are in line with the defined data architecture and/or with the goals of the business. Typically, a data architect maintains the metadata registry, oversees data management, optimizes databases and/or all data sources and more. Data architects are usually skilled at logical data modeling, physical data modeling, data policies development, data strategy, data warehousing, data querying languages and identifying and selecting a system that is best for addressing data storage, retrieval and management.

Additional Resources: http://stage.web.techopedia.com/definition/29452/data-architect

Data harmonization

Data harmonization is the process of bringing together data of different formats, metadata, and structure, often from different sources, and combining it so as to provide users with the ability to query or view it. *Additional Resources:* http://www.nationalacademies.org/hmd/~/media/Files/Activity%20Files/Quality/VSRT/Data-Harmonization/VSRT-WIB-DataHarmonization.pdf; https://www.icpsr.umich.edu/icpsrweb/content/DSDR/harmonization.html

Data management

Processes that include acquiring, validating, storing, protecting, and processing data to ensure accessibility, reliability, and timeliness for users.

Additional Resources: https://bioportal.bioontology.org/ontologies/MESH?p=classes&conceptid=D000079803

Data model

Representation of standards recognized to represent information, including official expression and structure of elements as well as the relationship between elements, to retain the expected meaning.

Additional Resources: https://bioportal.bioontology.org/ontologies/NCIT?p=classes&conceptid=http%3A%2F%2Fncicb.nci.nih.gov%

Data mapping

Creating relationships between similar data elements from different data models. This process can include connecting an item or symbol to a code or concept.

Additional Resources: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C142485

Data standardization

Data standardization uses a Common Data Model (*CDM*) to convert disparate datasets into the same structure. This allows for greater collaboration and interoperability when working with data.

The terms harmonization and standardization employed to attain the same goal of achieving data homogeneity, but while the standardization is about conformity, the harmonization is about consistency.

6.6.4 F

Federated query

A query that retrieves data from multiple, possibly geographically disparate databases (ex: SHRINE). Federated search is an information retrieval technology that allows the simultaneous search of multiple searchable resources. A user makes a single query request which is distributed to the search engines, databases or other query engines participating in the federation. The federated search then aggregates the results that are received from the search engines for presentation to the user.

Additional Resources: https://en.wikipedia.org/wiki/Federated_search

FHIR

Fast Healthcare Interoperability Resources (FHIR) Standard

FHIR is a standard for exchanging healthcare information electronically.

6.6.5 I

i2b2

*(See "***Common Data Model (CDM)***")*

6.6.6 O

OHDSI Network

OHDSI stands for Observational Health Data Sciences and Informatics and is pronounced as "Odyssey". It is a public initiative to enable analysis and sharing of real world health data (or observational data) between different institutes and companies. OHDSI represents a global network of research centers for standard analytical methods, tools and the OMOP

Common Data Model (CDM)

and vocabulary.

OMOP

*(See "***Common Data Model (CDM)***")*

Ontology An ontology is a specified model of a conceptualization containing concepts and the relationships between them.

An ontology allows data to be combined from multiple sources and integrated semantically.

- Example of an ontology: Gene Ontology (http://geneontology.org/) which is a computational model of biological systems. It contains genes, organisms, and other concepts from biology and the relationships that exist between them.

Note: The terms 'ontology', 'terminology', 'thesaurus', 'controlled language' and 'taxonomy' are used interchangeably. Often, you need to look deeper into what is being referenced to understand what is being used and why. An ontology for a data structure may contain elements that use concepts from a terminology or controlled vocabulary (for instance, a *data model* that contains the element "Disease" might have that element populated by terms from the "Disease Ontology" - which in this case, is actually a terminology.)

Additional Resources: http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C21270; doi:10.1016/j.datak.2015.11.00

PCORI

Patient-Centered Outcomes Research Institute

PCORNet

*(See "***Common Data Model (CDM)***")*

6.6.7 S

SCM (Sustainability and Change Management team)

The CD2H Sustainability and Change Management Task Team (SCM) of the CD2H Next Generation Data Sharing Core aims to aggregate, develop, curate, and disseminate content and tools that help encourage and guide organizations' understanding of and investments in *Data Harmonization* efforts.

Semantics Semantics in *Data Harmonization*: Semantics can be implicit within the structure of *data models*. One piece of metadata may have unstated relationships to other pieces of metadata, but they gain meaning from their inclusion in the same *data model* that would be absent would the metadata be viewed outside the model (think of a good example, maybe from a stem cell model?) This applies to both the structure of the *data model* itself, and any ontologies or vocabularies that are used. (another example using an ontology term)

Miriam Webster: The meaning or relationship of meanings of a sign or set of signs, especially connotative meaning. (https://www.merriam-webster.com/dictionary/semantics)

- Connote: to convey in addition to exact explicit meaning

'Semantic harmonization then is the process of combining multiple sources and representations of data into a form where items of data share meaning' (Cunningham, James & Speybroeck, Michel & Kalra, Dipak & Verbeeck, Rudi. (2017). Nine Principles of Semantic Harmonization. AMIA Annual Symposium Proceedings. 2016. 451-459)

Semantic Arts: Semantics is the study of meaning. By creating a common understanding of the meaning of things, semantics helps us better understand each other. Common meaning helps people understand each other despite different experiences or points of view. Common meaning helps computer systems more accurately interpret what people mean. Common meaning enables disparate IT systems – data sources and applications – to interface more efficiently and productively. (https://www.semanticarts.com/semantic-ontology-the-basics/)

6.6.8 T

Terminology

A terminology normalizes language, incorporating synonyms and alternate spellings in order to promote accurate discovery.

NCIT: A standardized, finite set of concepts, designations, and relationships for a specialized subject area. (http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C142730)

MeSH: Work consisting of lists of the technical terms or expressions used in a specific field. These lists may or may not be formally adopted or sanctioned by usage. (http://purl.bioontology.org/ontology/MESH/D020502)

- Example of a terminology: MeSH (https://meshb.nlm.nih.gov/search)

TriNetX(

https://www.trinetx.com/

)

TriNetX is a global health research network connecting healthcare organizations (including 35 *CTSA* Program hubs), biopharma and contract research organizations. The TriNetX platform enables cohort identification and hypothesis generation based on clinical data that can currently be sourced from a common data model *i2b2*, OMOP, NAACCR, etc.), flat files, or via natural language processing (NLP) of narrative documents.

6.7 References

"Harmonisation - DAPA Measurement Toolkit - Medical . . . ." https://dapa-toolkit.mrc.ac.uk/concepts/harmonisation. Accessed 27 Feb. 2020.

Hilal Atasoy, Brad N. Greenwood, Jeffrey Scott McCullough.The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. Annual Review of Public Health 2019 40:1, 487-500

"van Panhuis, W.G., Paul, P., Emerson, C. et al. A systematic review of barriers to data sharing in public health. BMC Public Health 14, 1144 (2014). https://doi.org/10.1186/1471-2458-14-1144

Geneviève LD, Martani A, Mallet MC, Wangmo T, Elger BS (2019) Factors influencing harmonized health data collection, sharing and linkage in Denmark and Switzerland: A systematic review. PLoS ONE 14(12): e0226015.

Jiang G, Solbrig HR, Iberson-Hurst D, Kush RD, Chute CG. A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki. Summit Transl Bioinform. 2010 Mar 1;2010:11-5. PMID: 21347136; PMCID: PMC3041544.

Marta Benet et al. Integrating Clinical and Epidemiologic Data on Allergic Diseases Across Birth Cohorts: A Harmonization Study in the Mechanisms of the Development of Allergy Project, *American Journal of Epidemiology*, Volume 188, Issue 2, February 2019, Pages 408–417, https://doi.org/10.1093/aje/kwy242

Knosp, Boyd M., William K. Barnett, Nicholas R. Anderson, and Peter J. Embi. "Research It Maturity Models for Academic Health Centers: Early Development and Initial Evaluation." Journal of Clinical and Translational Science 2, no. 5 (2018): 289-94. https://doi.org/10.1017/cts.2018.339

Knosp B, Craven CK, Dorr D, Campion T. Understanding enterprise data warehouses to support clinical and translational research. J Am Med Infom Assoc. Accepted for publication 2020 April 27. [Cited 2020 April 30]

Nicholson, J., Ioannidis, J. Conform and be funded. Nature 492, 34–36 (2012). https://doi.org/10.1038/492034a

Opinion: Expansion fever and soft money plague the biomedical research enterprise Henry R. Bourne, PNAS August 28, 2018 115 (35) 8647-8651; https://doi.org/10.1073/pnas.1813115115;

Rescuing US biomedical research from its systemic flaws, Bruce Alberts, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus, PNAS April 22, 2014 111 (16) 5773-5777; first published April 14, 2014 https://doi.org/10.1073/pnas.1404402111

Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data model considerations for clinical effectiveness researchers. *Medical care*, *50 Suppl*(0), S60–S67. https://doi.org/10.1097/MLR.0b013e318259bff4

Klann JG, Joss MAH, Embree K, Murphy SN (2019) Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLOS ONE 14(2): e0212463. https://doi.org/10.1371/journal.pone.0212463

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inf Assoc. 2010;17: 124–130. pmid:20190053

Visweswaran S, Becich MJ, D'Itri VS, Sendro ER, MacFadden D, Anderson NR, et al. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. JAMIA Open. pmid:30474072

Kimball R, Ross M. The data warehouse toolkit: the complete guide to dimensional modeling [Internet]. John Wiley & Sons; 2011. Available: https://books.google.com/books?

Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216: 574–578. pmid:26262116

Klann JG, Joss MAH, Embree K, Murphy SN, Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model, PLOS ONE, Feb 2019 https://doi.org/10.1371/journal.pone.0212463

Small Business (SBIR & STTR) Applications Directed at the Adoption of the Fast Healthcare Interoperability Resources (FHIR®) Standard. Notice Number: NOT-OD-19-127. Release Date: July 30, 2019

Fast Healthcare Interoperability Resources (FHIR®) Standard. Notice Number: NOT-OD-19-122. (Release Date: July 30, 2019)

CHAPTER

SEVEN

CHAPTER 7: REPOSITORY ARCHITECTURE FOR DATA DISCOVERY

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 Informatics playbook 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

EIGHT

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 Informatics playbook 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

NINE

CHAPTER 9: BEST PRACTICES OF ANNOTATING CLINICAL TEXTS FOR INFORMATION EXTRACTION TASKS

9.1 Intended Audience

Individuals who are leading or plan to lead the curation of gold standard annotation corpora for clinical information extraction tasks.

9.2 Current Version

Version: 0.2

Date: 1/26/2022

9.3 Status and Feedback Mechanisms

This is a work-in-progress chapter. We welcome your contribution and feedback via Github Issue.

9.4 Why is this important?

Unlike many of the structured data for informatics research, automatically removing protected health information from unstructured narratives according to the HIPAA safe harbor guideline is very challenging. Therefore, collecting clinical texts with human annotation from multiple institutions for collaborative research can become tricky, as the common quality assurance measures can only be taken after the clinical texts can be deidentified and manually reviewed. Since the manual annotation of information extraction tasks is typically very labor-intensive and has to be done by subject matter experts who may have limited bandwidth, any data quality issue that requires an update and re-review of the annotated data should better be addressed in the planning phase.

9.5 Roles

Similar to other manual data annotation tasks, an annotation task for information extraction may include several of the key contributors:

- **Principal Investigator (PI)**: solely responsible for the completion of the design, execution and dissemination with assistance from the rest of the team;
- **Project Manager (PM)**: overseeing the execution of the study and is in charge of the communications and content sharing within the team as well as the external stakeholders; monitoring the project progress to make sure the milestones and deliverables are at goal
- **Technical Lead (TL)**: preparing the original data and annotation schema; setting up the annotation system for the annotators to work on; making updates based on the feedback from the rest of the team
- **Information Technology Support (IT)**: working with the TL on the extract, transform and load (ETL) process to prepare the data, providing the platform (hardware, software or Platform-as-a-Service solutions)
- **Annotators (ANN)**: domain experts who receive directions from the PI and PM and conduct the generation of the annotations, either manually or with minor systematic assistance
- **Adjudicator**: the senior annotator(s) who can make the final decisions on the annotation discrepancies between the ANNs in double-annotation practices. Ideally, the adjudicator should not be an annotator to avoid conflict of interest.
- **Data Analyst (DA)**: run necessary benchmarks (e.g. inter-annotator agreement, numbers of annotations curated) to ensure quality

Please note that the list below provides only a general division of the roles and functions needed. Practically, it is very common to have one individual taking more than one role (e.g. the PI or TL also acting as the PM, the TL also acting partially as IT), which is acceptable as long as there is no conflict of interest (e.g. Annotators vs. Adjudicator regarding judgments for subjective annotations, Annotators vs. PM regarding progress monitoring).

9.6 Project Lifecycle

A common information annotation project life cycle includes the following steps

9.7 Takeaways

9.8 Examples

- Annotation guideline - Chronic Pain: [download]
- Annotation Guideline - Delirium: [download]
- Annotation guideline and algorithm - Fall occurrence: [code]

9.9 Open-sourced text annotation tools

- Brat: brat rapid annotation tool [link]
- Anafora: https://github.com/weitechen/anafora
- MAE: Multi-document Annotation Environment [link]
- MedTator: A Serverless Text Annotation Tool for Corpus Development [link]
- PubTator Central: PubTator Central (PTC) is a Web-based system providing automatic annotations of biomedical concepts such as genes and mutations in PubMed abstracts and PMC full-text articles. [link]

9.10 Annotation toolkits

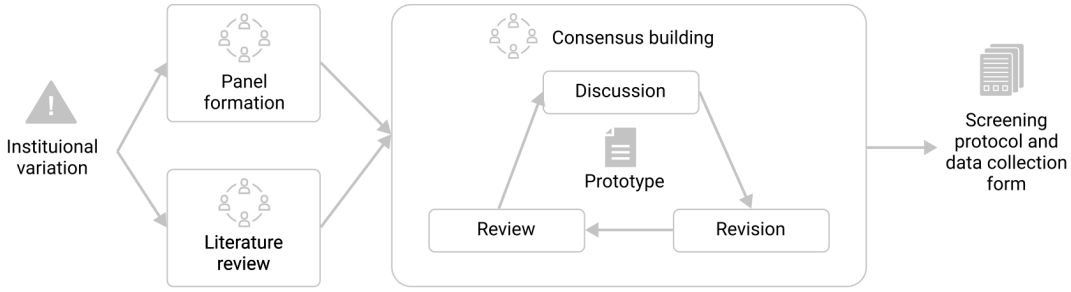
The following files should be opened by Microsoft Word.

- Standard Operating Procedure - Annotation - docx
- Instructions of Annotation Guideline Creation - docx
- Annotation Guideline Template - docx
- Checklist - docx

9.11 TRUST: clinical Text Retrieval and Use towards Scientific rigor and Transparent process.

9.11.1 Protocol Development

1. Protocol Development



Recommendations for Data Quality

Instituitional variation

- Compare clinical guideline, protocol, and definition

Human factor and coginitive bias

- Consensus development
- Rigor discussion about the choose of measurements and data collection points

Recommendations for Process Documentation

Study design

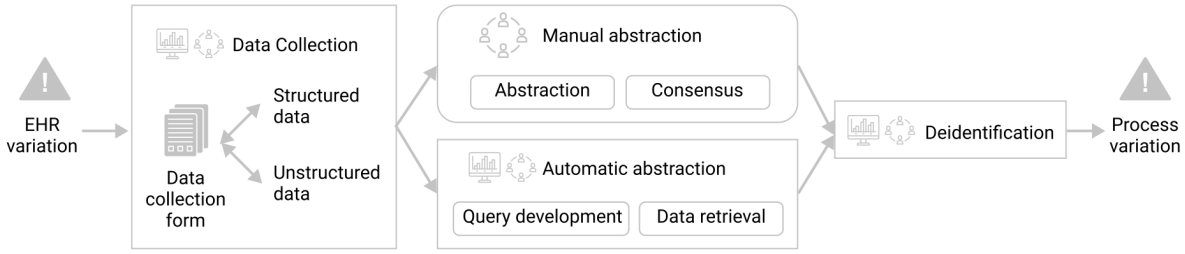
- Prospective disclosure of study plans, timing, and rationale for modifications

Screening protocol

- Clinical definitions of the study cohort (inclusion and exclusion criteria.) E.g., patients with clinically evident stroke any time before or up to 30 dyas after the imaning exam
- EHR definitions of the study cohort. E.g., ICD-9/10, CPT, search keyworkds

9.11.2 Data Collection

2. Data Collection



Recommendations for Data Quality

EHR system variation

- Compare data type, document structure, and metadata
- Conduct a semi-structured interview to obtain information about the context of use
- Data quality assessment, e.g., chart review, cross source examination

Process variation

- Standardize ETL process, data linkage methods, data identifiers

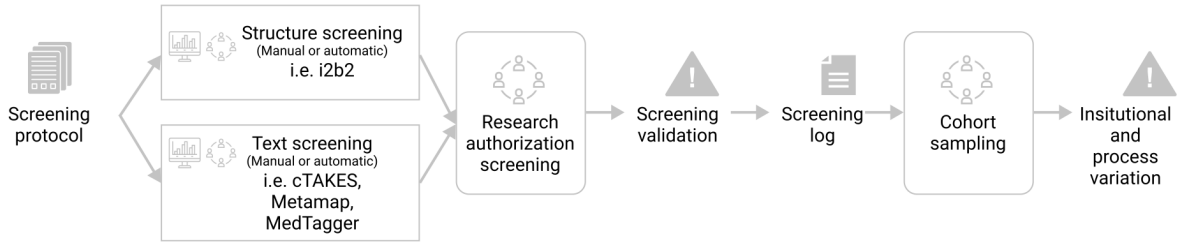
Recommendations for Process Documentation

Data collection form

- Data identifiers, e.g., document id, document date, and patient clinic number
- Data definitions, e.g., how is data measured, are there any composite definitions
- Metadata of ETL: process logs and additional information about when the data is created, moved, modified, or filtered
- Abstraction method: steward, tools (e.g., REDCap), and data collection methodologies (e.g., API, SQL, standardized tools)

9.11.3 Cohort Screening

3. Cohort screening



Recommendations for Data Quality

Institutional variation

- Calculate the number of eligible patients divided by screening population
- Calculate the ratio of the proportion of the persons with the disease over the proportion with the exposure

Process variation

- Use SOP to standardize screening process

Recommendations for Process Documentation

Screening protocol

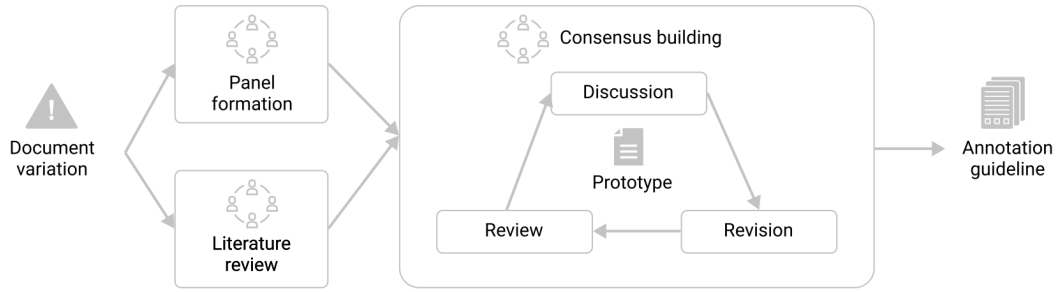
- Screening methods: e.g., manual or automatic
- Screening tools, e.g., i2b2, SQL
- Screening scripts

Screening log

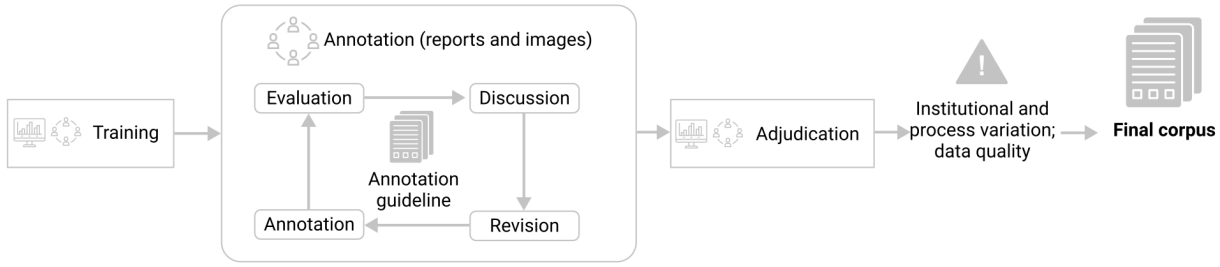
- Statistics about pre- and post-screening participants

9.11.4 Corpus Annotation

4. Guideline development



5. Corpus annotation



Recommendations for Data Quality

Document variation

- Compare document metadata, e.g., practice settings, sections
- Adoption of standards, e.g., HL7 CDA
- Compare the cosine similarity between two documents represented by vectors
- Conduct a sub-language analysis to assess syntactic variation

Human factor and cognitive bias

- Training
- Consensus development
- Calculate the degree of agreement among abstractors

Recommendations for Process Documentation

Annotation guideline: background of the study, annotation tool, annotation instructions, concept definitions, meeting notes for consensus development and team discussion, annotation examples, reference. More details can be found at: https://github.com/OHNLP/annotation-best-practices

9.11.5 Tips and Caveats

- All the digital contents (e.g. guideline drafts, schema, ETL scripts, IAA calculation scripts) should be version-controlled.

9.11.6 Communities

- BioNLP
- ClinicalNLP: 2019, 2020
  - Resources
- Health NLP 2018, 2019, 2020
- BioCreative/OHNLP 2018
- n2c2 NLP Research Data Sets
  - "Unstructured notes from the Research Patient Data Repository at Partners Healthcare."

9.12 Acknowledgment

9.12.1 Contributors to this playbook chapter

Name|Site|ORCID

Sijia Liu|Mayo Clinic|0000-0001-9763-1164

Sunyang Fu|Mayo Clinic|0000-0003-1691-5179

Hongfang Liu|Mayo Clinic|0000-0003-2570-3741

9.12.2 About the authors and contributors

This National Center for Data to Health (CD2H) playbook chapter is created on behalf of the Open Health Natural Language Processing Collaborative (https://github.com/OHNLP). Part of the work is also done through National COVID Cohort Collaborative (N3C) (https://ncats.nih.gov/n3c) Natural Language Processing (NLP) Subgroup under the Collaborative Analytics Workstream (https://covid.cd2h.org/N3C_analytics). More information can be found at https://github.com/OHNLP/N3C-NLP-Documentation/wiki.

9.12.3 Funding

Research reported in this playbook chapter was supported by the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health under award number U01TR002062. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

9.12.4 Resources

More about clinical information extraction research: awesome-clinical-nlp - OHNLP

9.13 References

- Sunyang Fu, TRUST: Clinical Text Retrieval and Use towards Scientific Rigor and Transparent Process, 2021/12, University of Minnesota
- Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan;77:34-49. doi: 10.1016/j.jbi.2017.11.011. Epub 2017 Nov 21. PMID: 29162496; PMCID: PMC5771858.
- Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K.J., Shen, F., Wang, L., Wang, Y., Wen, A. and Zhao, Y., Clinical Concept Extraction: a Methodology Review. Journal of biomedical informatics, p.103526.
- Liu S, Wen A, Wang L, et al. An Open Natural Language Processing Development Framework for EHR-based Clinical Research: A case demonstration using the National COVID Cohort Collaborative (N3C). arXiv; 2021.
- Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, Liu H. Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR Med Inform. 2021 Jan 27;9(1):e24008. doi: 10.2196/24008. PMID: 33502329; PMCID: PMC7875692.
- Mowery, D.L., Velupillai, S., South, B.R., Christensen, L.M., Martínez, D., Kelly, L., Goeuriot, L., Elhadad, N., Pradhan, S., Savova, G.K., & Chapman, W.W. (2013). Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. CLEF.
- Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019 Jul 2;47(W1):W587-W593. doi: 10.1093/nar/gkz389. PMID: 31114887; PMCID: PMC6602571.
- Wei-Te Chen, Will Styler. 2013. Anafora: A Web-based General Purpose Annotation Tool, In Proceedings of the NAACL-HLT, Companion Volume: Demonstrations, Atlanta, GA, USA, pp. 433-438.
- Kyeongmin Rim. MAE2: Portable Annotation Tool for General Natural Language Use. In Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, Portorož, Slovenia, May 28, 2016.
- He H, Fu S, Wang L, Liu S, Wen A, Liu H. MedTator: a serverless annotation tool for corpus development. Bioinformatics. 2022 Jan 4:btab880. doi: 10.1093/bioinformatics/btab880. PMID: 34983060.
- Carlson LA, Jeffery MM, Fu S, He H, McCoy RG, Wang Y, Hooten WM, St Sauver J, Liu H, Fan J. Characterizing Chronic Pain Episodes in Clinical Text at Two Health Care Systems: Comprehensive Annotation and Corpus Analysis. JMIR medical informatics. 2020;8(11):e18659.
- Fu S, Leung LY, Raulli AO, Kallmes DF, Kinsman KA, Nelson KB, Clark MS, Luetmer PH, Kingsbury PR, Kent DM, Liu H. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC medical informatics and decision making. 2020 Dec;20(1):1-2.

CHAPTER

TEN

CHAPTER 10: SELECTING AN ECONSENT PLATFORM

10.1 Authors

Ashley Clayton, MS | Sage Bionetworks | https://orcid.org/0000-0002-4570-2706

Christine Suver, PhD, PMP | Sage Bionetworks | https://orcid.org/0000-0002-2986-385X

Anita Walden, PhD| University of Colorado | https://orcid.org/0000-0002-3327-7423

Chunlei Wu, PhD | The Scripps Research Institute | https://orcid.org/0000-0002-2629-6124

Marie Rape, RN, BSN | University of North Carolina at Chapel Hill | https://orcid.org/0000-0002-0608-7093

10.2 Intended audience

Individuals interested in the selection, implementation, and use of electronic informed consent (eConsent) platforms in clinical research settings. The framework presented here may also be adapted to help select other platforms and tools.

10.3 Key Words

eConsent; informed consent; electronic consent; Technology Assessment Tools; Technology Evaluation

10.4 Version history

Chapter version: 1.0

Date: 21 Oct 2022

10.4.1 eConsent Assessment Framework Tools

V1.0 Qualtrics implementation: eConsent Needs Assessment Tool (Qualtrics)

V1.0 PDF rendition: eConsent Needs Assessment Tool (PDF)

V1.0 Qualtric implementation: eConsent Evaluation Tool (Qualtrics)

V1.0 PDF rendition: eConsent Evaluation Tool (PDF)

10.5 Why is this important?

Selecting or developing a suitable eConsent platform can be challenging and may be informed by a variety of factors at both the study and organizational level. Furthermore, there is a spectrum of eConsent features, from the simple rendition of an informed consent form in an electronic format to a multimedia interactive eConsent experience in different languages. Prior to selecting an eConsent platform, researchers must:

1. assess their eConsent needs from research, administrative, technical, and financial standpoints, as well as the needs of consenting individuals.
2. evaluate platforms to determine which ones meet their needs.

This may require input from consenting individuals, study teams, information technology (IT) experts, regulatory personnel, institutional decision makers, and other relevant stakeholders. This chapter describes several considerations for the selection of an eConsent platform for clinical research studies, using a novel assessment framework to evaluate needs and eConsent platform features. It should also be noted that while this chapter focuses on eConsent platforms, the considerations and related tools for eConsent selection may be transferable to the assessment of other platform types and research tools (e.g., electronic data capture tools).

10.6 Development of an eConsent Assessment Framework

Choosing any tool or product requires balancing needs, wants, finances, and constraints with the technology tool or product specification. Similarly, when selecting an eConsent platform to implement in biomedical research, one must understand what the platform is meant to achieve, how it will be used, and by whom. In collaboration with partners at the National Center for Data to Health (CD2H), representatives of the North Carolina Translational and Clinical Sciences (NC TraCS) Institute at the University of North Carolina (UNC) surveyed Clinical and Translational Science Awards (CTSA) program hubs about their eConsent needs, experiences, and satisfaction. The information was used by Sage Bionetworks to develop and pilot an **eConsent Assessment Framework** composed of two related tools:

- **eConsent Needs Assessment Tool** (i.e. "what do I need in an eConsent platform?")
- **eConsent Evaluation Tool** (i.e. "what features does eConsent platform 'x' have?")

10.7 Understanding Your eConsent Needs - Performing a Needs Assessment

Several factors may inform the decision to select or build an eConsent platform for clinical research use. Ultimately, the selected platform should meet as many anticipated requirements as possible, based on an assessment of eConsent needs, wants, and constraints at the user, study, and organizational levels. Answering the following questions using the **eConsent Needs Assessment Tool** can help recognize what matters most in selecting an eConsent platform:

10.7.1 Who should evaluate the organization's or projects' eConsent needs?

Consider which study team member(s) and organizational role(s) are best suited to determine the eConsent needs for a specific study, a group of related studies, and/or the enterprise level. Those individuals assessing needs must be able to differentiate essential features that are 'required' from those that are 'preferred' and 'not important'. The selection of these individuals may be informed by their familiarity with informed consent, eConsent practices and regulatory practices, prior experience with clinical research studies, technical expertise, and organizational structure, among other factors.

Example stakeholders to identify study or organization eConsent requirements

10.7.2 What are the study-specific vs. enterprise-level eConsent platform needs?

Consider whether the eConsent platform would be used only for a specific study or for multiple studies organization-wide. At the study-specific level, eConsent platform needs may be informed by factors such as:

- regulatory and ethical compliance (e.g., domestic vs. international studies),
- study population (e.g. familiarity with technology),
- study type (e.g., pediatric studies), or
- budgetary considerations (e.g., free-to-use vs. license-based platforms), among other factors.

At the organization level, eConsent platform needs may be informed by:

- the projected number of studies over a period of time,
- the need for integration with existing infrastructure (e.g., Clinical Trials Management System (CTMS), Clinical Data Management System (CDMS), current organization-wide licenses and subscriptions), and
- implementation requirements (e.g., security review and validation for new platforms), among other factors.

10.7.3 How will the eConsent platform be used?

eConsent platform needs may be informed by the intended user audience and anticipated user setting. Consider what elements (e.g., features) of an eConsent platform are important and/or required to support ease of use and cultural competency among study team members and organizational personnel that will be using or interacting with the platform. The level of existing training materials, anticipated user support, and available resources to support both implementation and maintenance of the platform should also be considered in the assessment of platform needs. Additionally, requirements related to ease of use and engagement from the perspective of consenting individuals should be considered. eConsent platform requirements may also be informed by the setting(s) under which the consent is anticipated to be performed (e.g., in-person vs. remote), or use of specific equipment (e.g., laptops vs. tablets), among other factors.

10.7.4 What eConsent platform features are required?

Existing free-to-use and license-based platforms offer a range of features and levels of customization. Some features may be required (i.e., essential) while others may be preferred (i.e., "nice-to-have") but not required for a given study, or "not important." General categories of platform features and customization include, but are not limited to:

- Language
- Accessibility
- Automation
- Reporting and metrics

- Security
- Integration

Note:

Understanding eConsent Platform Needs:

- eConsent Needs Assessment Tool: Developed by Sage Bionetworks in collaboration with CD2H and North Carolina Translational and Clinical Sciences (NC TraCS) Institute at the University of North Carolina (UNC). One of two tools within the eConsent Assessment Framework, this tool can be used by research teams and organization stakeholders to assess their eConsent platform needs and requirements, exploring each of the needs-based considerations described in this chapter.
- Readings:
  - Doerr M, Moore S, Suver C. Elements of Informed Consent. In: Sage Bionetworks [Internet]. [cited 30 June 2022]. Available here.
  - Cobb NL, Edwards DF, Chin EM, Lah JJ, Goldstein FC, Manzanares CM, Suver CM. From paper to screen: regulatory and operational considerations for modernizing the informed consent process. J Clin Transl Sci. 2022 Mar 28;6(1):e71. doi: 10.1017/cts.2022.379. PMID: 35836789; PMCID: PMC9257776.
  - Moore S, Tassé AM, Thorogood A, Winship I, Zawati M, Doerr M. Consent Processes for Mobile App Mediated Research: Systematic Review. JMIR Mhealth Uhealth. 2017 Aug 30;5(8):e126. doi: 10.2196/mhealth.7014. PMID: 28855147; PMCID: PMC5597795.

10.8 Evaluating eConsent platforms for potential implementation

eConsent platforms available today offer a range of features and levels of customization,with varying cost, regulatory compliance, implementation ease, and security measures). This can make it challenging for prospective users to assess which platform(s) may be best suited for implementation based on their particular needs and requirements. Additionally, the expertise to adequately evaluate eConsent platforms against user-level, study-level, and organizational-level requirements may reside across multiple individuals and roles.

The **eConsent Evaluation Tool** provides a set of harmonized criteria for objective evaluation of eConsent platform characteristics and feature availability. This approach facilitates the comparison and potential selection of eConsent platforms. The tool is structured to evaluate a range of feature categories by representatives across different roles and areas of expertise within an organization (e.g., study team members, regulatory experts, IT professionals), allowing for consolidated, multi-user evaluation of a particular eConsent platform.

Note:

Evaluating eConsent platforms

- eConsent Evaluation Tool: Developed by Sage Bionetworks in collaboration with CD2H and North Carolina Translational and Clinical Sciences (NC TraCS) Institute at the University of North Carolina (UNC). One of two tools within the eConsent Assessment Framework, this tool provides a set of criteria for research teams to harmonize their evaluation of eConsent platform characteristics and feature availability, facilitating ease of comparison and potential selection of eConsent platforms.

10.9 Use and applicability of the eConsent Assessment Framework

The eConsent Assessment Framework was designed to facilitate the selection of eConsent platforms in clinical research settings. Although various sources of information on informed consent and eConsent practices exist, this framework, consisting of an **eConsent Needs Assessment Tool** and an **eConsent Evaluation Tool**, may be helpful to reconcile specific eConsent needs against the availability of eConsent platform features. Existing familiarity with both general study participant consent and eConsent practices may be helpful for adoption of these tools.

Similarly to first identifying the eConsent needs, acknowledging the involvement and expertise to adequately evaluate eConsent platform features may be distributed across a range of study team personnel and organizational roles, it should be noted that the **eConsent Evaluation Tool** can be completed by several personnel in different roles within an organization (e.g., study team members, regulatory experts, IT professionals), to provide a consolidated, multi-user evaluation of a given eConsent platform. While the framework was developed to facilitate the selection of eConsent platforms, a similar assessment framework structure developed with relevant stakeholder and community input could be utilized to assess other platforms and research tools for potential implementation (e.g., electronic data capture tools).

10.10 Limitations of the eConsent Assessment Framework

Several limitations were identified and acknowledged throughout the refinement and launch of the eConsent Assessment Framework, and have been described below. These limitations may also be applicable in the evaluation of other research-related platforms and applications.

- **Language:** The language of both tools is quite technical in nature. Future iterations of these tools should adjust for a wider range of technical literacy to improve accessibility, user experience, and potential generalizability to other applications.
- **Roles:** The questions included in both tools have been prioritized and organized based on the following themes and categories of anticipated user roles: Study Team (e.g., investigator), IT (e.g., information security officer), and Regulatory / Compliance (e.g., data privacy officer). However, the scope of users across the research community, domains of expertise, and areas for evaluation with respect to eConsent platforms may extend beyond these three roles. Future iterations of this tool should accommodate an expanded set of user roles accordingly.
- **Barriers to platform evaluation:** The assessment of certain platform features using this framework may be limited and/or prohibited by proprietary software or confidential information protections, thus preventing a comprehensive evaluation of some eConsent platforms. Some evaluation questions may be more appropriately directed to the vendor of an eConsent platform, as some information may not be accessible to the general research community to adequately evaluate certain features.
- **Access to evaluation data:** At this time, completed assessments of eConsent needs and platforms are accessible only to study teams performing these assessments. However, the evaluation and potential comparison of eConsent needs across larger, unrelated study portfolios, at the broader organizational level, and across organizations may be of interest to the research community. Access to a comparison matrix of eConsent platform evaluation data across multiple users - similar to a consumer report - may be helpful to facilitate the comparison of different eConsent platforms. Additionally, with the ongoing development and evolution of eConsent platforms, access to updated and/or versioned evaluations of a particular platform will need to be considered.

10.11 Summary

- **Different requirements will inform the selection of an eConsent platform for use in a single/small group of studies versus in multiple studies across anorganization or enterprise.**
- **First, identify the stakeholder needs for an eConsent platform (eConsent Needs Assessment Tool)**
- **Determine the current and long term future needs, such as if the eConsent platform will be used for FDA-regulated studies, or the number and types of studies.**
- **Identify multiple eConsent platforms of interest to evaluate.**
- **Evaluate multiple platforms using an eConsent evaluation list of required and desired features (eConsent Evaluation Tool).**
- **Involve a variety of stakeholders consisting of different roles, including potential study participants, study team and regulatory administrators, and IT/security officers in the eConsent needs assessment process and platform evaluation.**
- **The eConsent Assessment Framework can be adapted and applied to other applications**

10.12 About the authors and contributors to this playbook chapter

The National Center for Data to Health (CD2H) is committed to the development of tools for assessing and comparing best practice approaches in research across the Clinical and Translational Science Awards (CTSA) program hubs and broader scientific community (see also https://cd2h.org/maturity). The CD2H tasked Sage Bionetworks with the development and iterative refinement of an eConsent Assessment Framework for use by research teams to help them evaluate, select, or build an eConsent tool or platform. This process informed several of the considerations described in this chapter regarding the evaluation and selection of eConsent platforms.

10.13 Acknowledgements

The authors would like to thank Xindi Guo and Emily Lang at Sage Bionetworks and representatives at NC TraCS Institute (University of North Carolina) for their contributions to the development of the assessment questions and piloting the eConsent Assessment Framework. Thank you also to representatives at the following institutions for piloting the assessment framework: Sage Bionetworks, University of Colorado, The Scripps Research Institute, Emory University, University of Wisconsin, and West Virginia University.

10.14 Funding

This work was supported by the National Center for Data to Health (CD2H) of the National Institutes of Health (NIH) under award number U24TR002306 and by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through award number UL1TR002489. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

CHAPTER

ELEVEN

TUTORIAL: HOW TO WRITE A CHAPTER USING MARKDOWN

This document demonstrates which formats and text styling we provide as well as best practices, how-to's, dos and don'ts and more.

11.1 Headings

All 6 headings are available, but you should only use `Heading 1` for the chapter title, and use `Heading 2`, `Heading 3`, ... for section, subsections, ...

- Example:

```
# Heading 1 for Chapter title

## Heading 2 for Section

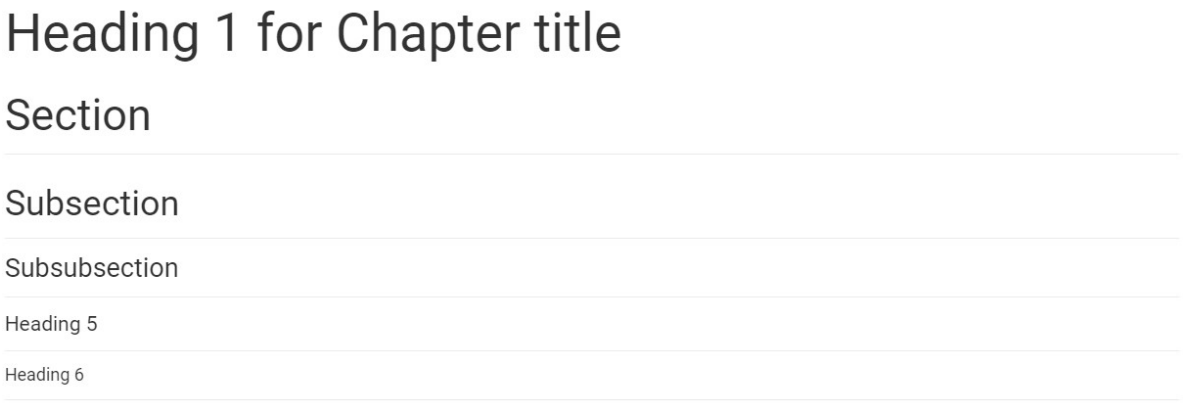
### Heading 3 for Subsection

#### Heading 4 for Subsubsection

##### Heading 5

###### Heading 6
```

It should render like this:



11.2 Emphasis

You can use bolds, italics and a mix of them. To add them select your text then click one of these buttons (Bolds, italics)

- Example:

```
**Bold**

*Italic*

~~Strikethroughs~~

***Mixes (bold and italic)***
```

It should render like this:

Bold

Italic

~~Strikethroughs~~

Mixes (bold and italic)

11.3 Lists

Nest lists as much you like. NOTE: numbered lists are not supported yet but will be converted to bulleted lists. To add a deeper list add a tab at the start of the line.

- **Unordered example**:

```
* Item 1
* Item 2
  * Item 2a
  * Item 2b
```

It should render like this:

- Item 1
- Item 2
  - Item 2a
  - Item 2b

- **Ordered example**:

```
1. Item 1
1. Item 2
1. Item 3
   1. Item 3a
   1. Item 3b
```

It should render like this:

1. Item 1
2. Item 2

3. Item 3
    1. Item 3a
    2. Item 3b

11.4 Links

The embeded links will be automatically converted to clickable links. You can also use the markdown syntax `[link text](link url)` to create a link to any URL, including the link to any section or subsection.

- Example:

```
https://cd2h.org/
[CD2H website](https://cd2h.org/)
Links to [the "Code block" section](#code-block)
```

It should render like this:

https://cd2h.org/

CD2H website

Links to *the "Code block" section*

11.5 Tables

Tables are also supported. 1st row will be bolded, use this to explain what your data is in each column.

- Example

```
First Header | Second Header
------------ | -------------
Content from cell 1 | Content from cell 2
Content in the first column | Content in the second column
Content in the first column | Content in the second column
```

It should render like this:

11.6 Figures

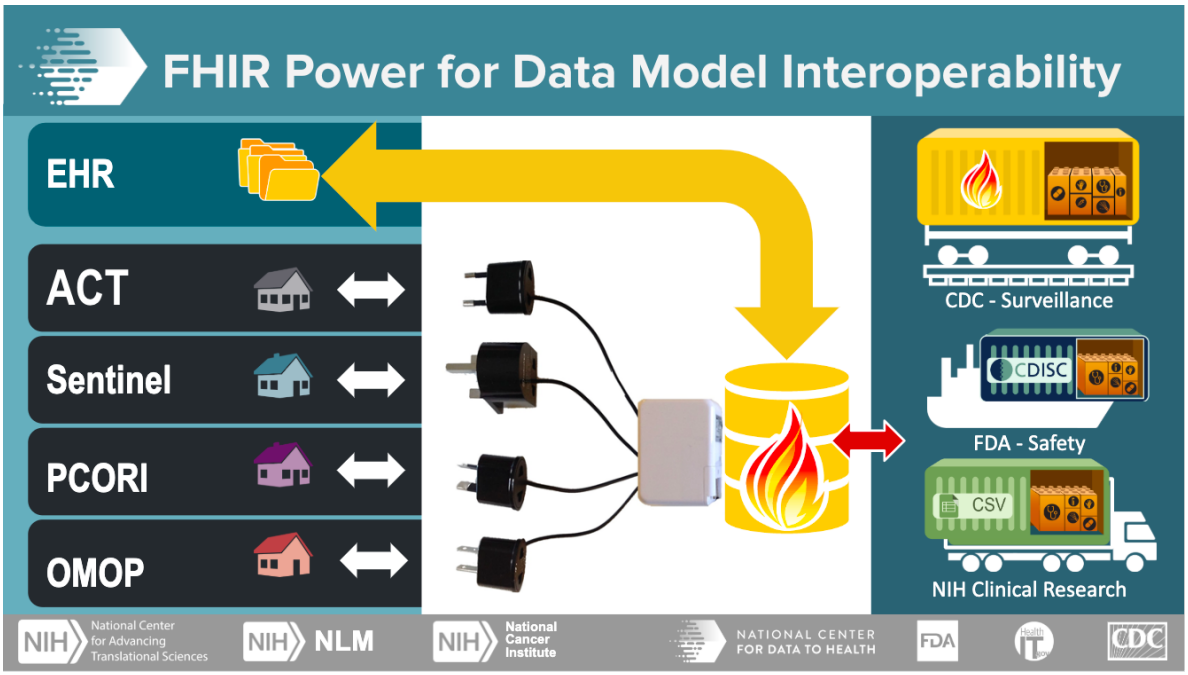
You can add figures as you like but each figure will take a row of each own.

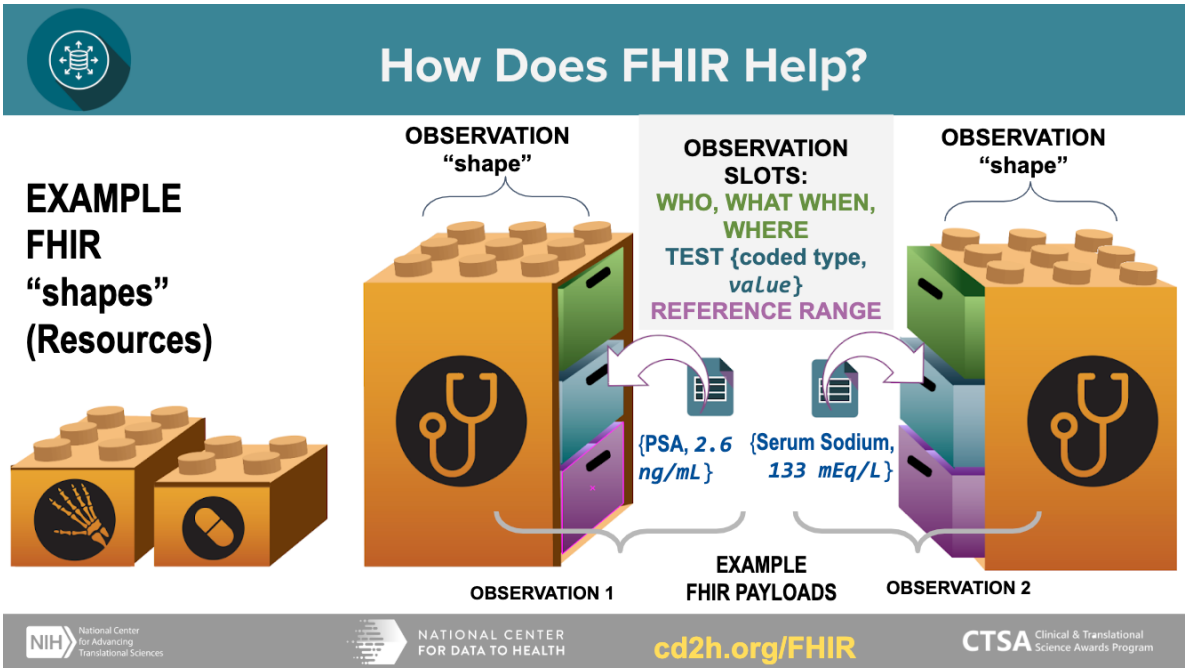
- Example

```
![demo figure 1](../_static/img/chapter_6_fhir_1.jpg)

![demo figure 2](../_static/img/chapter_6_fhir_2.jpg)
```

It should render like this:


FHIR Power for Data Model Interoperability
EHR
ACT
Sentinel
PCORI
OMOP
CDC - Surveillance
CDISC
FDA - Safety
CSV
NIH Clinical Research
National Center for Advancing Translational Sciences
NLM
National Cancer Institute
NATIONAL CENTER FOR DATA TO HEALTH
FDA
CDC


How Does FHIR Help?
EXAMPLE FHIR "shapes" (Resources)
OBSERVATION "shape"
OBSERVATION SLOTS:
WHO, WHAT WHEN, WHERE
TEST {coded type, value}
REFERENCE RANGE
OBSERVATION "shape"
{PSA, 2.6 ng/mL}
{Serum Sodium, 133 mEq/L}
OBSERVATION 1
EXAMPLE FHIR PAYLOADS
OBSERVATION 2
National Center for Advancing Translational Sciences
NATIONAL CENTER FOR DATA TO HEALTH
cd2h.org/FHIR
CTSA Clinical & Translational Science Awards Program

11.7 Videos

To include a video, you can upload it to Youtube first and then add its YouTube URL like this:

- Example

```
```eval_rst
.. youtube:: https://www.youtube.com/watch?v=0JPjw1_iRKY
```
```

or just the Youtube id:

```
```eval_rst
.. youtube:: 0JPjw1_iRKY
```
```

It should render like this

https://youtu.be/0JPjw1_iRKY

11.8 Code block

You can embed a block of code in the text, with the optional syntax-highlighting as well. ``two backticks``

- Example

```
```
function fancyAlert(arg) {
  if(arg) {
    $.facebox({div:'#foo'})
  }
}
```
```

It should render like this:

```
function fancyAlert(arg) {
  if(arg) {
    $.facebox({div:'#foo'})
  }
}
```

Optionally, you can enable the syntax-highlighting:

```
```javascript
function fancyAlert(arg) {
  if(arg) {
    $.facebox({div:'#foo'})
  }
}
```
```

It should render like this:

```
function fancyAlert(arg) {
  if(arg) {
    $.facebox({div:'#foo'})
  }
}
```

You can replace `javascript` with other language types like `python`, `bash`, etc.

11.9 Inline code

You can also embed an inline code in a paragraph.

- Example

```
You can call this Python function `do_analysis` to get the result.
```

It should render like this:

You can call this Python function `do_analysis` to get the result.

11.10 Math formula

You can include math formula using double dollar signs.

- Example

```
$$\omega = d\phi / dt$$

$$I = \int \rho R^{2} dV$$
```

It should render like this:

$$\omega = d\phi / dt$$

$$I = \int \rho R^{2} dV$$

11.11 Special text box

You can add some special text box to emphasize some content.

- Example

```
```eval_rst
.. note::

  This is a special note.
```

```eval_rst
.. warning::

  This is a warning message.
```
```

(continues on next page)

(continued from previous page)

````
```eval_rst
.. hint::

  Here you can provide a hint message.
```
````

It should render like this:

Note: This is a special note.

Warning: This is a warning message.

Hint: Here you can provide a hint message.

11.12 Citations

To add a citation in the text, you can add a link where the citation is referenced:

```
Data modeling is the process of determining which data elements will be stored and␣
↪how they will be stored, including their relationships and constraints. The␣
↪structure and definitions of a data model define what data can be stored, how␣
↪values should be interpreted, and how easily data can be queried [1](#kix.
↪unntzb98ia8x).
```

And then add the actual citation text in the "**References**" section, usually the last section of the chapter:

```
## References

[](#kix.unntzb98ia8x) Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data model␣
↪considerations for clinical effectiveness researchers. *Medical care*, *50␣
↪Suppl*(0), S60–S67. [https://doi.org/10.1097/MLR.0b013e318259bff4](https://doi.org/
↪10.1097/MLR.0b013e318259bff4)
```

The inline citation link are associted to the corresponding reference item based on the same link, e.g. `#kix.unntzb98ia8x` in this example. This example should renders like this:

> Data modeling is the process of determining which data elements will be stored and how they will be stored, including their relationships and constraints. The structure and definitions of a data model define what data can be stored, how values should be interpreted, and how easily data can be queried *1*.

And you can cite the same reference multiple times in the text, as long as you use the same link. For example, this is the first citation *2* and we can cite it again here *2*.

11.13 A few extra notes

- Try not to mix "Emphasis" or "Link" elements with other elements such as headers and titles because it might add additional unwanted new lines.
- Try not to add too much formatting in your "Heading" elements like bolds, italics etc. Also try not to extend these to multiple lines.
- Images will span full width of the document.
- Direct videos are not available, but you can paste a youtube link and it will be embedded in the final result automatically.
- Don't use "Heading" elements for anything other than headers. Each heading will be added to the table of contents so long text should be avoided.
- Tables inside cells are not supported.
- All headings will automatically be added to the table of contents.

11.14 References

Kahn, M. G., Batson, D., & Schilling, L. M. (2012). Data model considerations for clinical effectiveness researchers. *Medical care*, *50 Suppl*(0), S60–S67. https://doi.org/10.1097/MLR.0b013e318259bff4

Knosp B, Craven CK, Dorr D, Campion T. Understanding enterprise data warehouses to support clinical and translational research. J Am Med Infom Assoc. Accepted for publication 2020 April 27. [Cited 2020 April 30]

CHAPTER

TWELVE

INDICES AND TABLES

- genindex
- modindex
- search