

Reproducible Workflows Biomedical Research

P11 2018
Berlin, Germany

Contributors

Leslie McIntosh	Research Data Alliance, U.S., Executive Director
Oya Beyan	Aachen University, Germany
Anthony Juehne	RDA, U.S., Sr. Implementation & Evaluation Analyst
John Graybeal	Stanford University, U.S.
Mark Musen	Stanford University, U.S.
Health Data IG	https://www.rd-alliance.org/groups/health-data.html

Disclosures

Leslie McIntosh - CEO

Anthony Juehne - Chief Science Officer

Ripeta, LLC - Co-Founders

Investment from Digital Science

What is the Issue?

Scientific Publication

vs.

Scientific Pipeline

The Concern

How do we ensure everything within the scientific compendium is fully transparent and appropriately accessible to achieve reproducibility?

Assumptions

1. Reproducible research is important
2. Process of sharing biomedical data not comprehensively documented
3. This work can be overwhelming

What is reproducibility?

Computation Reproducibility:

If we took your data and code/analysis scripts and re-ran it, we can reproduce the numbers/graphs in your paper

Empirical Reproducibility:

We have enough information to re-run the experiment the way it was originally conducted

Replicability (Results Reproducibility):

We use your exact methods and analysis, but collect new data, and we get the same results

RESEARCH DATA ALLIANCE



Previous Work

Presented at RDA P8 & P9

Funding Support

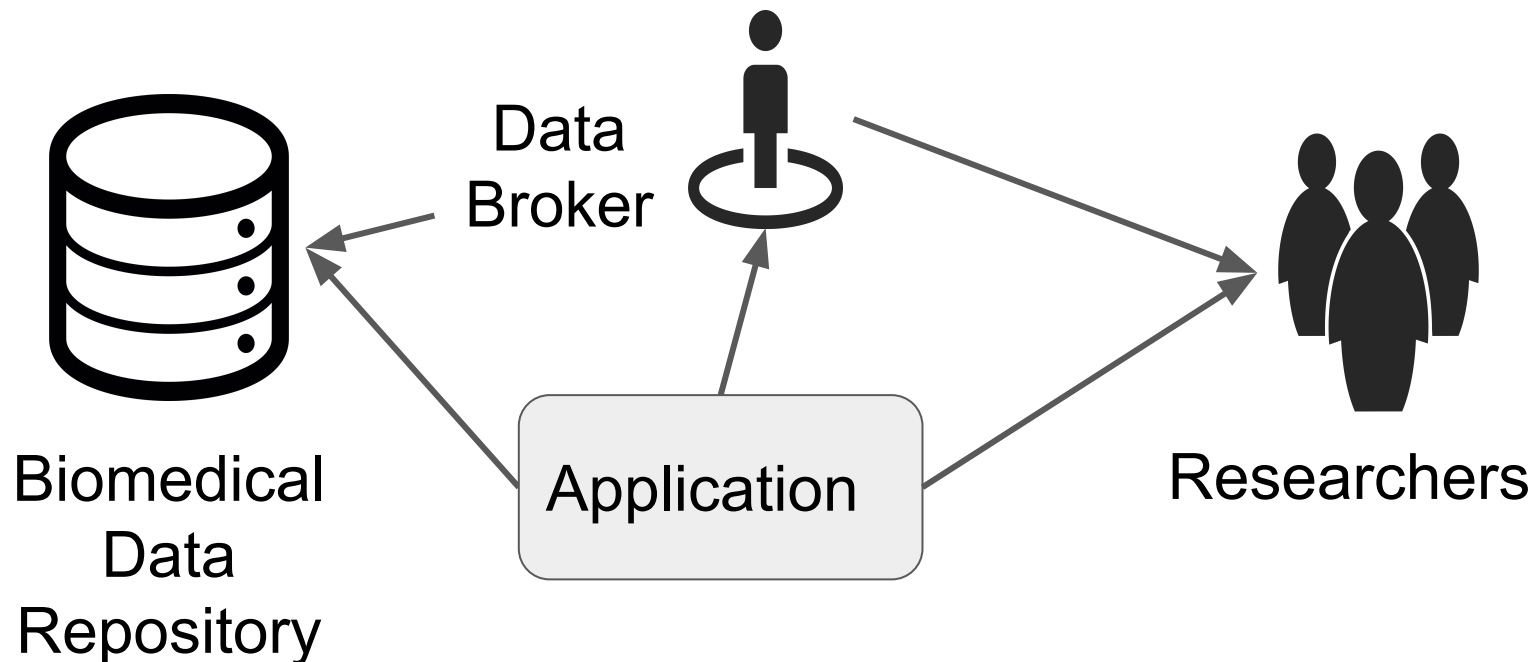
Washington University Institute of Clinical and Translational Sciences

NIH CTSA Grant Number UL1TR000448 and UL1TR000448-09S1

MacArthur Foundation 2016 Adoption Seeds program Foundation through a sub-contract with Research Data Alliance

BDaaS

Biomedical Data as a Service



Move some of the responsibility of
reproducibility

Biomedical Biomedical
Researcher  Pipeline

Framework for Biomedical EHR

Publication Overview and Bibliographic Information (21 items)

Is the research hypothesis-driven or
hypothesis-generating?

Hypothesis Driven
Hypothesis Generating
Unclear

Database and Data Collection (63 items)

Publication states database(s) source(s) of data?

Yes/No

.....
*Publication states database(s) source(s) of data in the
following location:

Not Stated
Supplementary materials
Appendix
Body of Text

.....
Query methodology

Manual extraction
Digital extraction through query interface
Digital extraction through honest broker
Not Applicable/Not Stated

Methods: Data Mining and Cleaning (19 items)

Does the research involve natural language processing or
text mining?

Yes/No

.....
*Is the text mining software application proprietary or
open?

1. Proprietary
2. Mixed
3. Open

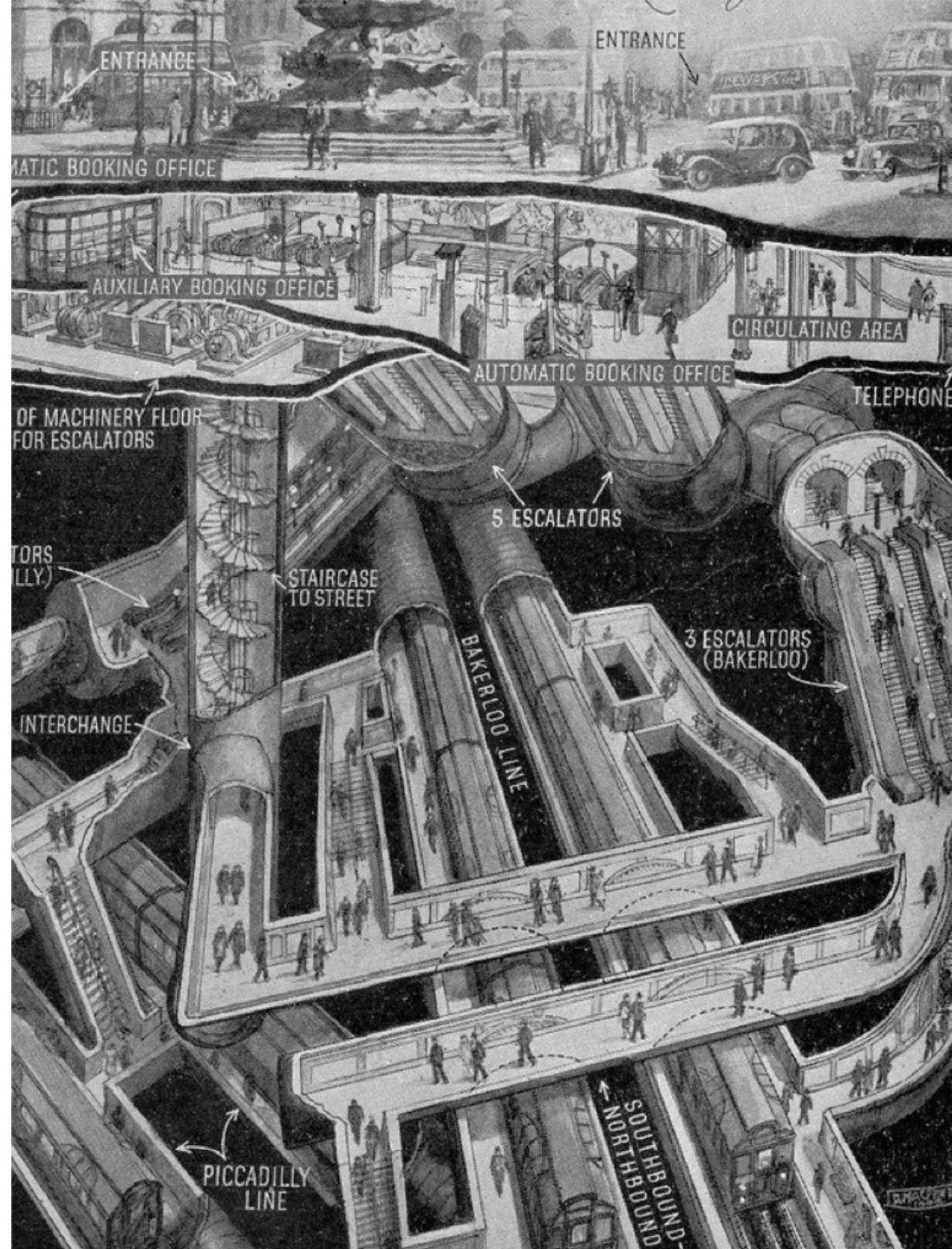
*If multiple applications were used, please select all
options that apply.*

Workflow for Reproducible Data Brokerage

Health Data IG Task Force P10

[Make progress
towards]

Develop a roadmap for
reproducible workflows
within biomedical
informatics data
brokerage

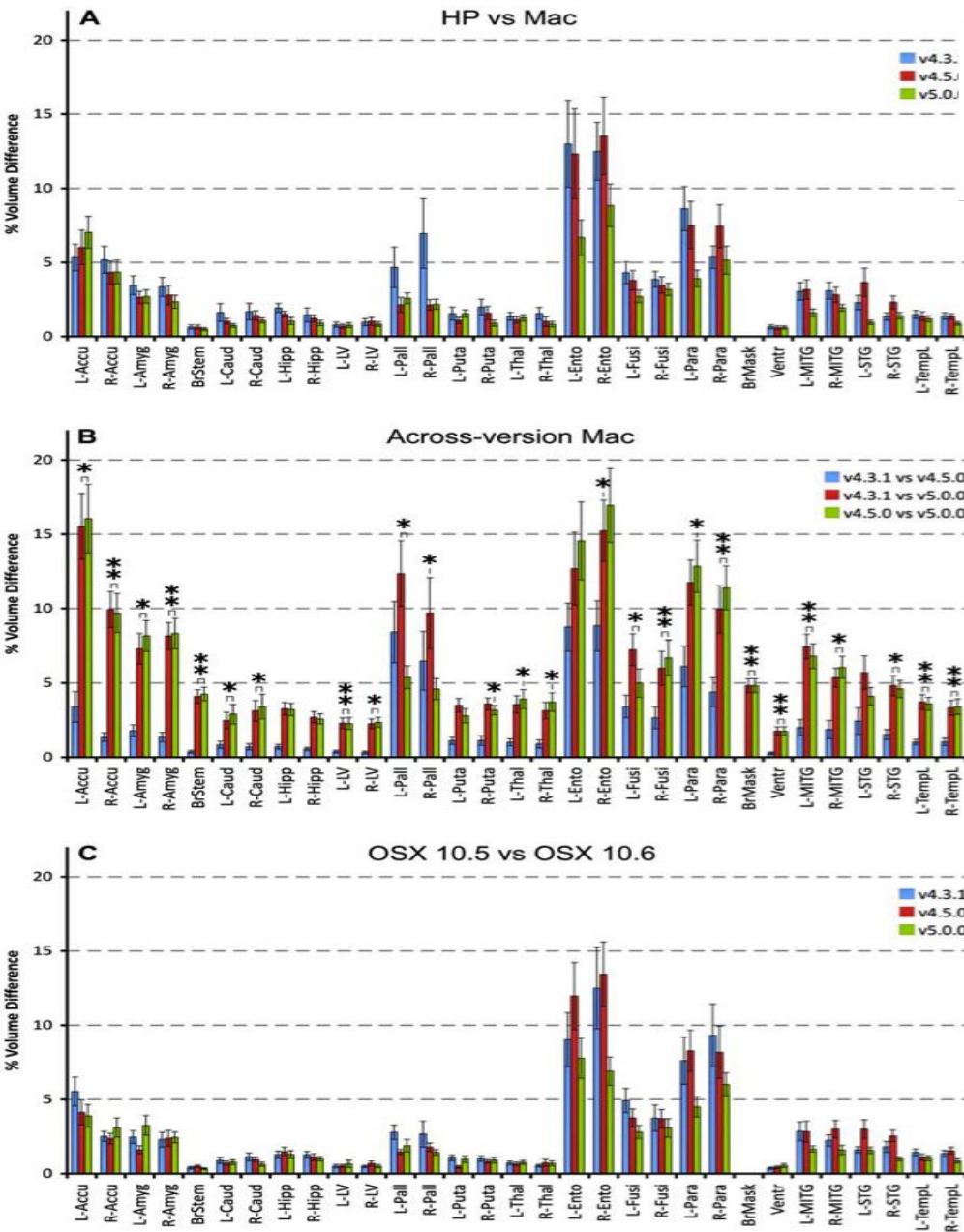


What will it take?

1. Enhance transparency & intelligibility of adopted methods
2. Improve capacities to confidently verify and validate results
3. Access materials essential to reproduce methods

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild^{1,2,*}, Petra Habets^{1,2}, Heidi I. L. Jacobs^{1,2,3}, Ron Mengelers^{1,2}, Nico Rozendaal^{1,2}, Jim van Os^{1,2,4}, Machteld Marcelis^{1,2}



What are the differences that make a difference?

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	x	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-		Applicability across different settings
Port	-	-	x	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	x		Apply code in different settings, Re-purpose
Independent x (orthogonal)							x	Sufficiency of information, independent verification

Reproducible Scientific Workflows

“Reproducibility implies repetition and thus a requirement to also move back – to **retrace** one’s steps, **question** or **change** assumptions, and move forward again.”

Millman, K. J., & Perez, F. (2014). Developing Open-Source Scientific Practice (V. Stodden, F. Leisch, & R. D. Peng, Eds.). In *Implementing Reproducible Research* (CRC the R series, pp. 149-183). Boca Raton, FL: Taylor & Francis Group, LLC.

The Why

“The construction of a
scientific heritage
where anyone can validate the work
of others and build upon it.”¹

Millman, K. J., & Perez, F. (2014). Developing Open-Source Scientific Practice (V. Stodden, F. Leisch, & R. D. Peng, Eds.). In *Implementing Reproducible Research* (CRC the R series, pp. 149-183). Boca Raton, FL: Taylor & Francis Group, LLC.

RDA Health Data IG P11

Challenges of Health Data Services

- Data curation processes
 - requires to interact with multiple systems
 - needs manual effort and crafting to map, transform, clean the data
 - typically queries and scripts are written case bases
 - ETL processes varies highly depending on the task and the curator

Challenges of Health Data Services

- Documenting metadata
 - very limited documentation if any
 - there is no recommendation / guideline on what to document
 - too time consuming
- How to capture the metadata
 - no guidance on what is useful (problem of granularity)
 - what are the available standards ?
- Sharing health data curation metadata
 - no standard way to find, access and interpret this metadata

Methodology

- Identify data curation processes
 - which activities are carried out at the different phases of the data services
- Document challenges related to activities
- Explore the available standard stack
 - RDA, W3C, ISO, research communities

Methodology

Outcomes:

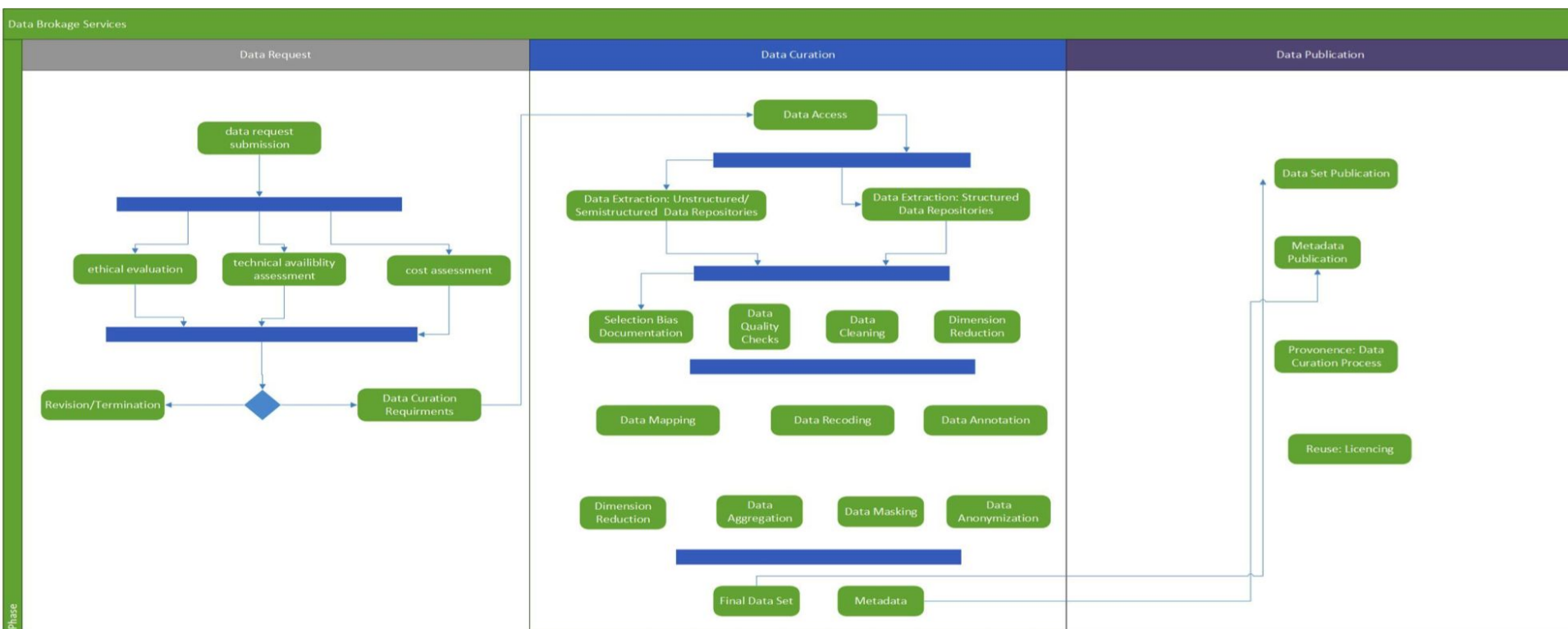
- Perform a Gap Analysis
 - relevance , maturity levels, ...
 - identify needs for further standards and methods
 - communicate the need with relevant groups
- Develop Adoption and Training Guidelines
 - Documentation of state-of-the-art methods and standards for clinical data curation
 - best practices for capturing and storing data curation metadata
 - identify use cases

Reproducible Workflows for Health Data Services Centers

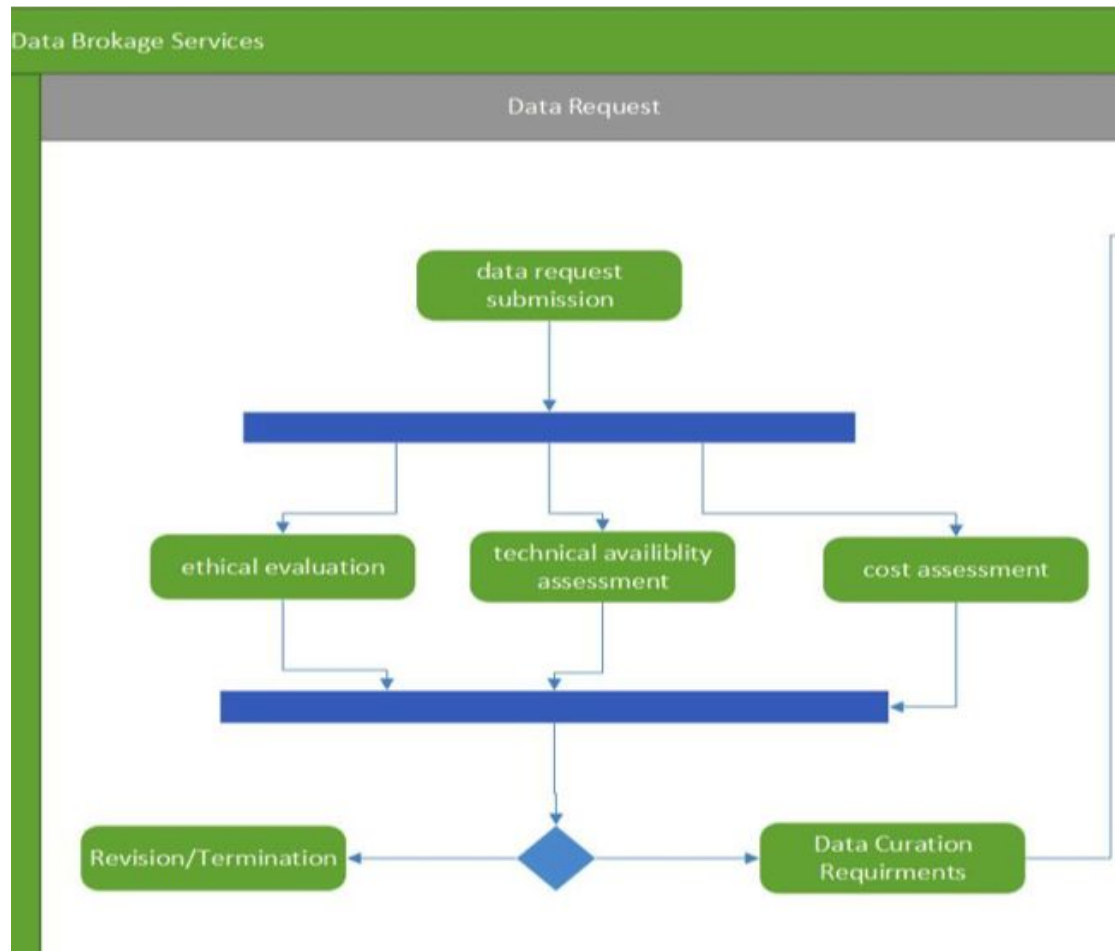
Data Request

Data Curation

Data Sharing



Data Request Elements



Ethical Assessment

- IRB protocol approval
- Consent form
- Approval date
- Reviewing institution

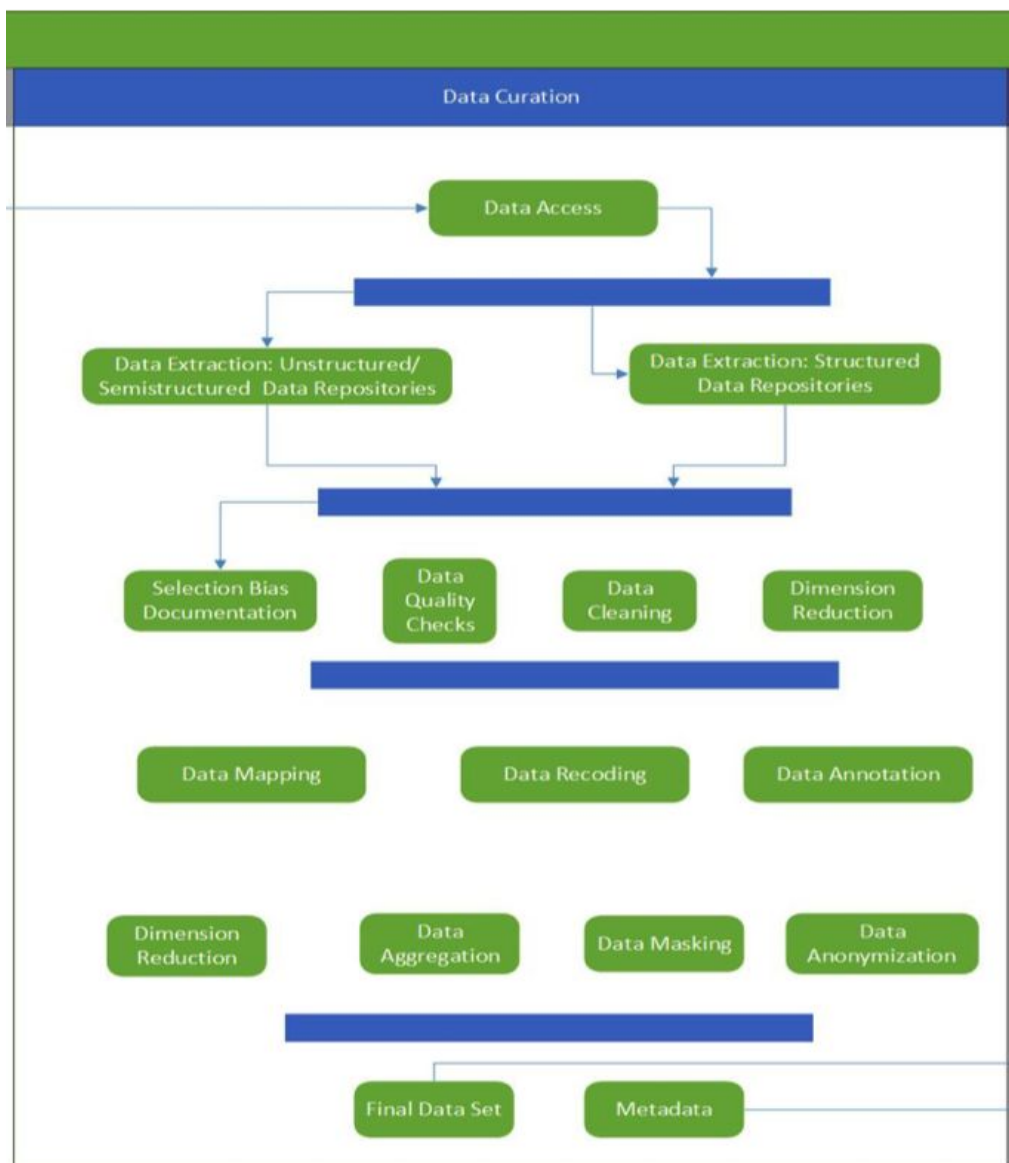
Cost Assessment

- Grant information
- Center expected hours of work

Feasibility Assessment

- Is the requested data available?
- Does it satisfy the clinical question?

Data Curation Processes



What is the source of data?

- DOI and Data citation?
- Multi-Site/Multi-registry Collection

Are query scripts FAIR?

- Interoperable
- Commented

How are participant inclusion and exclusion criteria operationalized?

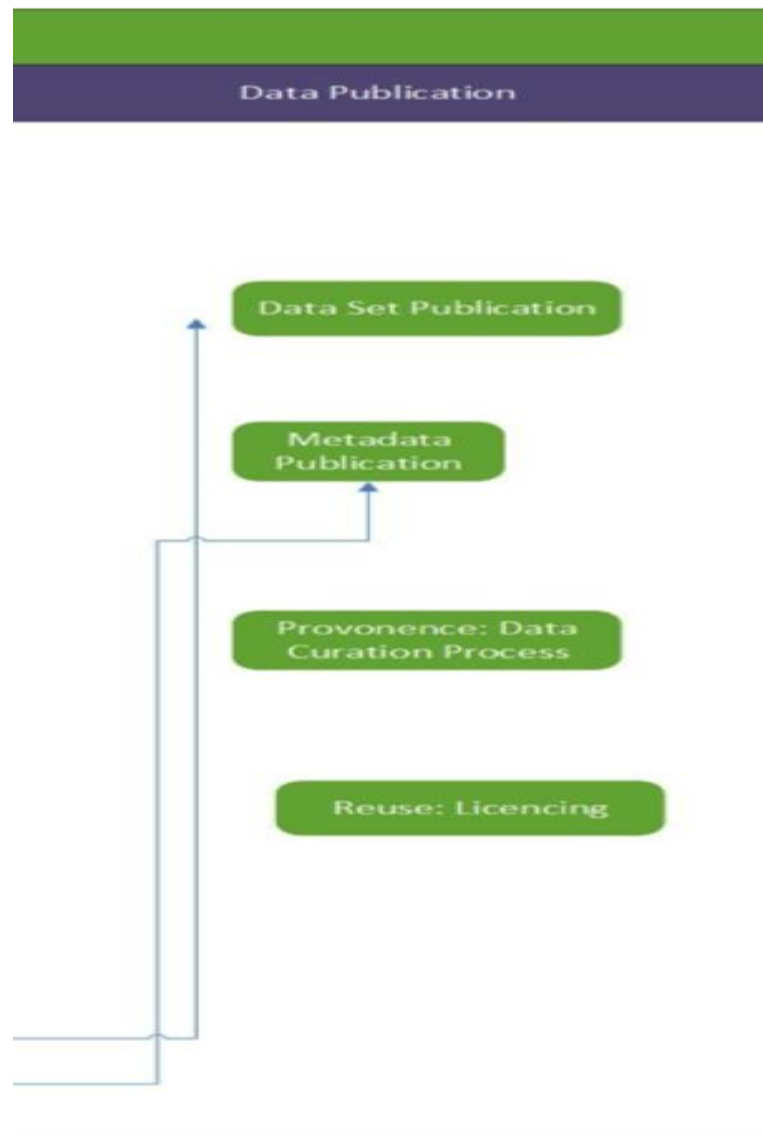
- Ontologies and standardized vocabs
- Limitations

How are data cleaned, mined, and merged?

- FAIR software code

Is there a FAIR final dataset or data dictionary?

Data Sharing and Publication



Can we apply FAIR principles to data products of the health data service workflows ?

- Rich metadata
- Persistent identifiers
- Licensing
- Common protocols to access

How to share data curation metadata together with the data?

Is it possible to have a minimum information reporting standard for data services?

How to make the data curation workflow metadata FAIR?

Draft Working Document

Current Health Data Service Center Workflow

https://docs.google.com/spreadsheets/d/1-uSocVpju4_fBcMDBgW2LxG3EpvBNRedOQuings2Cok/edit?usp=sharing

- Phases of Data Services
- Curation Activity
- Explanation
- Reproducibility Challenges
- Possible Metadata Formats
- Relevant Community Resources/Standards

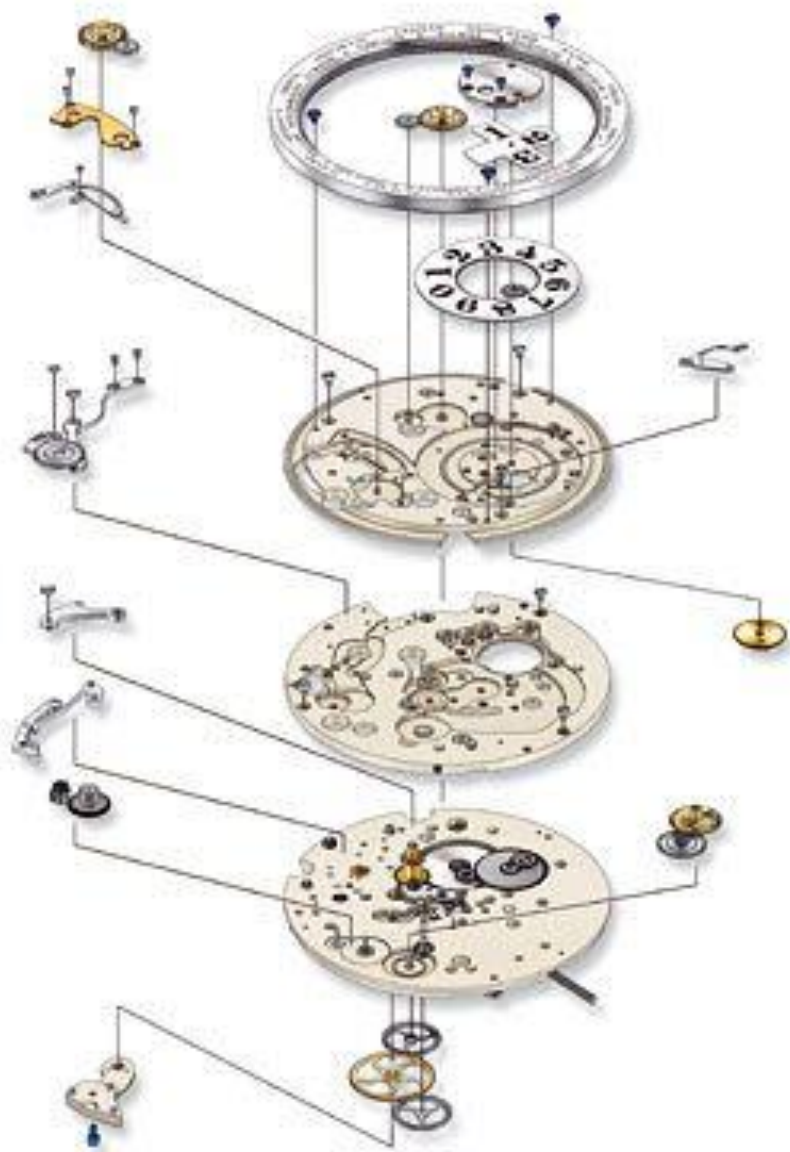
Stanford Adoption Case Study

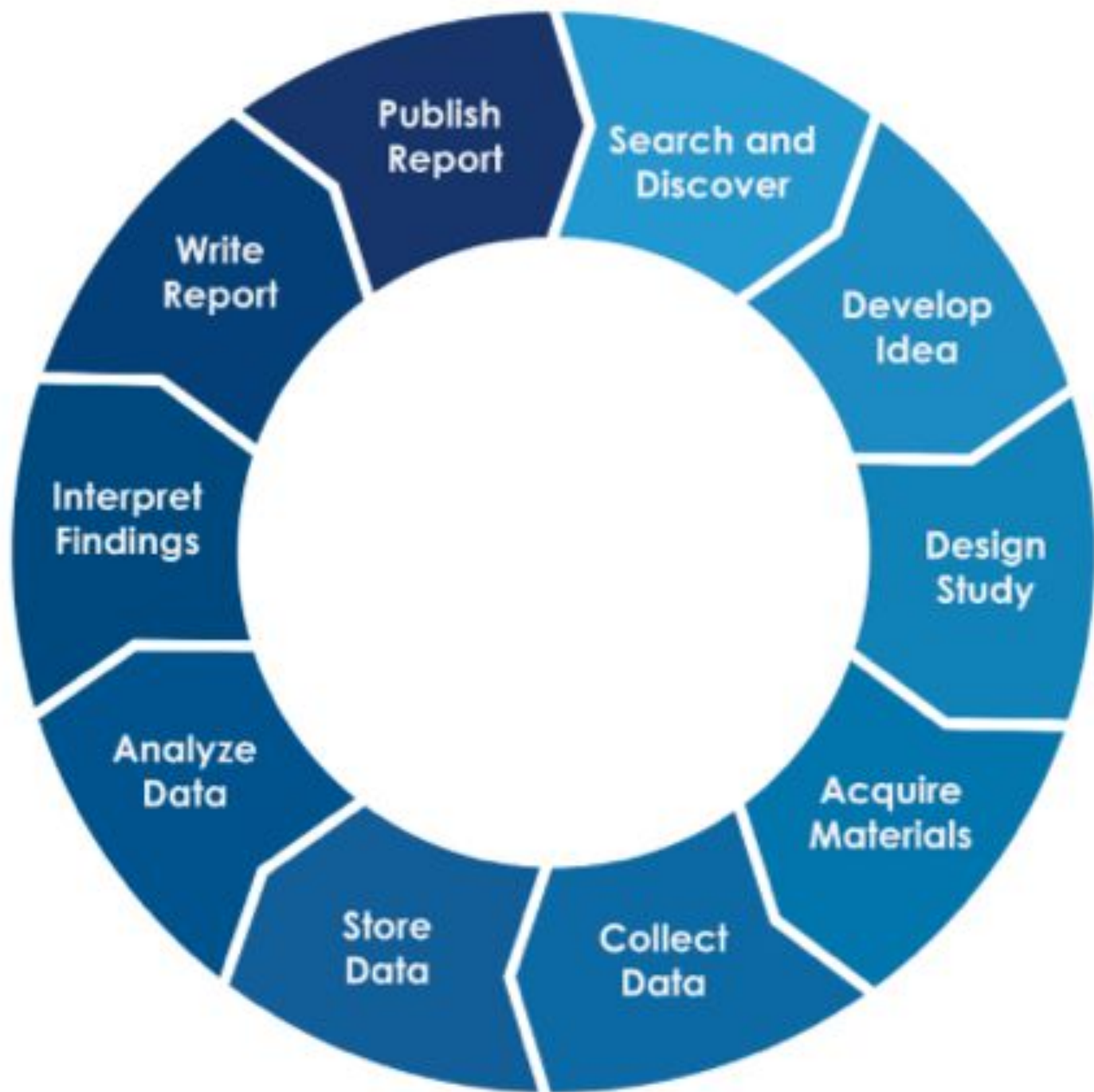
Scientific Publication

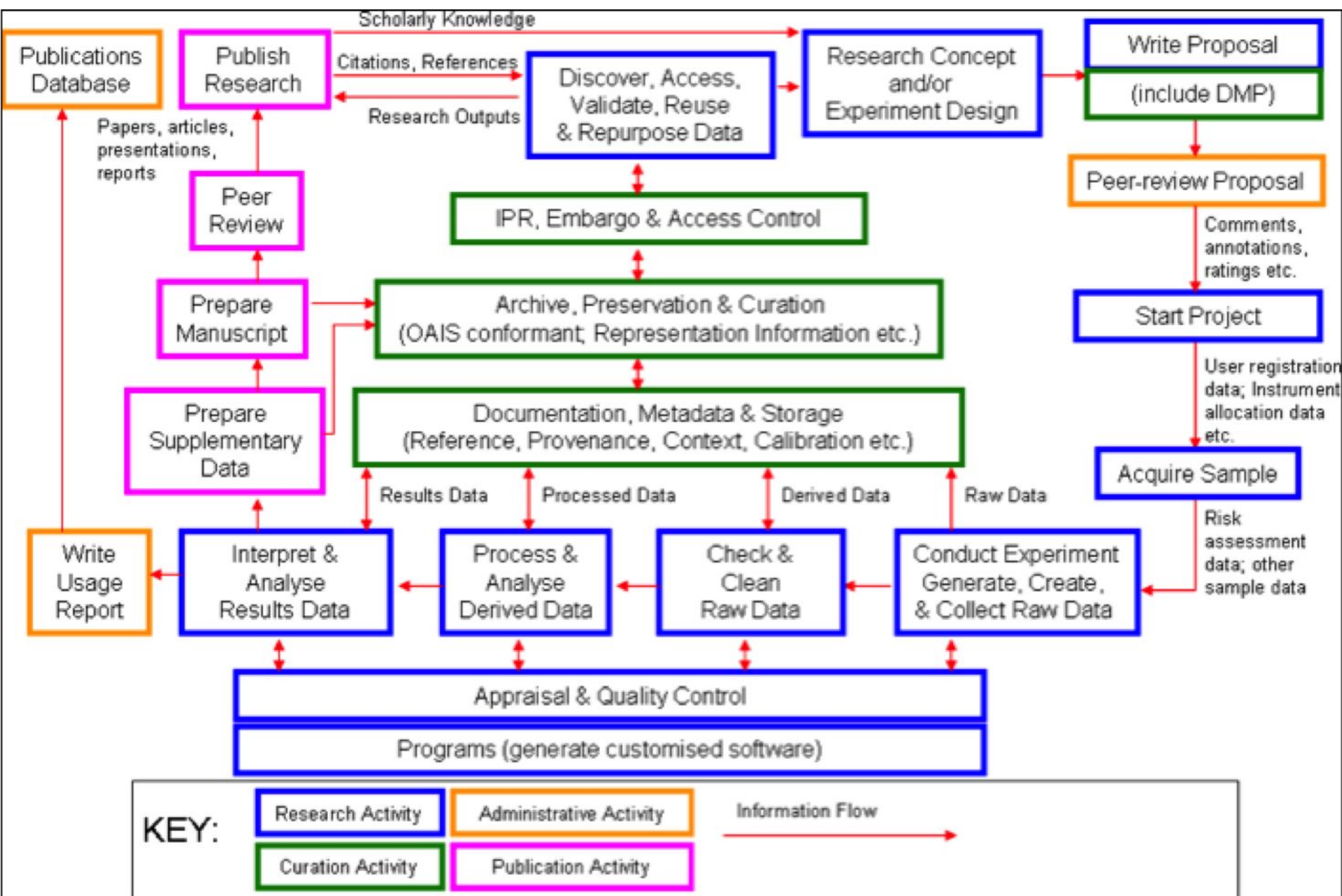


Scientific Workflow









	Title	Created
	Cost Assessment	1/9/18 2:15 AM
	Data Access	1/10/18 7:00 AM
	Data Extraction - Structured	1/10/18 6:13 PM
	Data Extraction_Unstructured Data	1/10/18 6:16 PM
	Ethical Assessment Review	1/9/18 1:50 AM
	Participant Exclusion Criteria	2/9/18 8:55 PM
	Participant Inclusion Criteria	1/11/18 2:32 AM
	Project Metadata	1/11/18 6:59 AM
	RDA_Health Data Service Center_Data Brokerage Workflow	1/11/18 7:21 AM

Types of Data

- Text and numeric descriptors
 - Names, titles, journal of publication...
- Software files (or associated links)
 - Query, cleaning, nlp...
- Data files (or associated links)
 - tabular, clinical notes, data dictionaries
- Identifiers (hopefully persistent)
 - DOIs, URLs, Grant IDs...
- Clinical Ontology Variables
 - Diagnosis, Procedure, Medication, Labs...

Bibliographic and Project Metadata



Project Metadata



Project Title



Project Principal Investigator



Institution of Project Principle Investigator

Health Data Service Center Principle Investigator or Director

Institution of Health Data Service Center

Ethical Assessment and Review



IRB Approving Institution



IRB Approval Date



IRB Protocol ID #



Link to IRB Protocol



Full Text of IRB Protocol

Data Access and Collection

▼ Data Access

- Database Name
- Database DOI
- Database URL
- Institution Hosting Database
- Institutional Department or Center Overseeing Database

▼ Data Extraction - Structured

- Query Script_Structured
Data_File Name
- Link to Shared Query
Script_Structured Data
- Query Script_Structured
Data_ File Format
- Query Script_Structured
Data_Software Language
- Query Script_Structured
Data_File Version #
- Query Script_Structured
Data_Author(s)
- Query_Structured
Data_Execution Date

→ Data Extraction_Unstructured Data

Learning Effective Treatment Pathways for Type-2 Diabetes from a clinical data warehouse

[Rohit Vashisht](#), PhD,¹ [Ken Jung](#), PhD,¹ and [Nigam Shah](#), MBBS, PHD¹



 rohit43 modified temp table names		Latest commit d34242a on Dec 12, 2017
 R	Update runT2DOutcomeStudy.R	4 months ago
 extras	second commit	11 months ago
 inst	modified temp table names	3 months ago
 man	updated R functions	11 months ago
 DESCRIPTION	Update DESCRIPTION	11 months ago
 DiabetesTxPath.Rproj	first commit	a year ago
 NAMESPACE	first commit	a year ago
 README.Rmd	Update README.Rmd	4 months ago
 README.html	updated R functions	11 months ago

Linking and describing cohort collection files

 **rohit43 / DiabetesTxPath**

 Watch **1**  Star **0**  Fork **3**

 Code  Issues **0**  Pull requests **0**  Projects **0**  Insights

Branch: master ▾ **DiabetesTxPath / R /**  Create new file  Find file  History

 **rohit43** Update runT2DOutcomeStudy.R Latest commit 93cd11c on Dec 2, 2017

..

 buildOutcomeCohort.R	Update buildOutcomeCohort.R	4 months ago
 createExposureCohorts.R	Update createExposureCohorts.R	4 months ago
 createNegativeControlOutcomeCohorts.R	Update createNegativeControlOutcomeCohorts.R	4 months ago

▼ Data Extraction - Structured

Query Script_Structured Data_File Name	createExposureCohorts.R
Link to Shared Query Script_Structured Data	https://github.com/rohit43/DiabetesTxPath/blob/master/R/createExposureCohorts.R
Query Script_Structured Data_File Format	.R
Query Script_Structured Data_Software Language	R, SQL, and JSON
Query Script_Structured Data_File Version #	2.0
Query Script_Structured	Rohit Vashisht, Jamie Weaver

Mapping Ontologies, Vocabularies, and Standards

Element Name	Element Description
 Participant Inclusion Criteria	Description of cohort inclusion and extraction criteria operationalized through data brok

 a 1 .. N 

Enter Field Title

ICD 9 Codes_Participant Inclusion

Enter Field Description (Help Text)

List of ICD 9 Codes Operationalizing cohort inclusion criteria

Enter Default Value

 VALUES

MULTIPLE

REQUIRED

SUGGESTIONS

HIDDEN

 INSTANCE TYPE

Name	Type	Source	Identifier	No. Values
DISEASES AND INJURIES	Branch	ICD9CM	001-999.99	-

SEARCH

→ Data Extraction_Unstructured Data

▼ Inclusion Criteria

- ICD 9 Codes_Participant Inclusion
- ICD10 Codes_Participant Inclusion
- HCFA/HCPCS Procedure Codes_Participant Inclusion
- CPT Codes_Participant Inclusion
- RXNORM Medication Codes_Participant Inclusion
- SNOMED Medication Codes_Participant Inclusion
- LOINC Laboratory Codes_Participant Inclusion

```
{"ConceptSets":[{"id":0,"name":"HbA1c_v2","expression":{"items":[{"concept":{"CONCEPT_ID":3004410,"CONCEPT_NAME":"Hemoglobin A1c (Glycated)","STANDARD_CONCEPT":"S","INVALID_REASON":"V","CONCEPT_CODE":"4548-4","DOMAIN_ID":"Measurement","VOCABULARY_ID":"LOINC","CONCEPT_CLASS_ID":"Lab Test","INVALID_REASON_CAPTION":"Valid","STANDARD_CONCEPT_CAPTION":"Standard"}}]}
```



LOINC Laboratory Codes_Participant Inclusion	
1	▼
2	▼
3	▼
4	▼
5	▼
6	▼
7	▼
8	▼
9	▼
10	▼


```
"Data Extraction_Unstructured Data": {
  "@context": {
    "Text Extraction Script_Unstructured Data_File Name": "https://schema.metadatacenter.org/properties/e05ead1d-4a0b-4000-8000-000000000000",
    "Text Extract Script_Unstructured Data_Software Language": "https://schema.metadatacenter.org/properties/719ec416-4000-8000-8000-000000000000",
    "Query Script_Unstructured Data_File Name": "https://schema.metadatacenter.org/properties/51645b80-69c3-4e51-9b94-000000000000",
    "Query Script_Unstructured Data_File Format": "https://schema.metadatacenter.org/properties/59f02b9f-ea6b-4ea5-b500-000000000000",
    "Query script_Unstructured Data_Software Language": "https://schema.metadatacenter.org/properties/3e9e8745-4b96-4000-8000-000000000000",
    "Query Script_Unstructured Data_Author(s)": "https://schema.metadatacenter.org/properties/ff32801e-7253-4a75-97d6-000000000000",
    "Query Script_Unstructured Data_Execution Date": "https://schema.metadatacenter.org/properties/2777b740-4ef4-4fac-0000-000000000000",
    "Text Mining Script_Unstructured Data_Author(s)": "https://schema.metadatacenter.org/properties/a221b43d-3832-4ed0-0000-000000000000",
    "Text Mining Script_Unstructured Data_Date Executed": "https://schema.metadatacenter.org/properties/4166f856-17ba-4000-8000-000000000000",
    "Link to Shared Query Script_Unstructured Data": "https://schema.metadatacenter.org/properties/6a4d573b-a83d-4e02-0000-000000000000",
    "Link to Share Text Mining Script_Unstructured Data": "https://schema.metadatacenter.org/properties/9865a4cd-b9e0-4000-8000-000000000000",
    "Query Script_Unstructured Data_File Version": "https://schema.metadatacenter.org/properties/4675ccfb-446f-4795-9000-000000000000",
    "Text Extraction Script_Unstructured Data_File Version": "https://schema.metadatacenter.org/properties/26c3f13a-e000-0000-8000-000000000000"
  },

```

```
"Inclusion Criteria": {
  "@context": {
    "ICD 9 Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/5931e35d-4ba9-41d2-a1f4-bbda63-000000000000",
    "CPT Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/1b2b6f54-7f92-4efb-87b9-555640b3-000000000000",
    "RXNORM Medication Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/36d6470a-fe30-482b-0000-000000000000",
    "SNOMED Medication Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/674e211e-a657-43d6-0000-000000000000",
    "LOINC Laboratory Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/08e85ded-623c-4fec-0000-000000000000",
    "HCFA/HCPCS Procedure Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/5af3c336c-5dc1-4000-8000-000000000000",
    "ICD10 Codes_Participant Inclusion": "https://schema.metadatacenter.org/properties/867ab168-46d9-48c5-8f59-941a11-000000000000"
  },
  "ICD 9 Codes_Participant Inclusion": [
    {}
  ],
  "CPT Codes_Participant Inclusion": [
    {}
  ],

```

Proposed Next Steps

- Include ontologies and vocabularies from OMOP model
- Assess feasibility with more use cases
- Iterative reviews with other members of the RDA
- Adjudicate feedback from AMIA
- Continued collaboration with CEDAR to test and develop functionality
- How can we make usability easier?

Potential Future Use Cases



Australian Government

Australian Digital Health Agency



OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



IHME Celebrating 10 years of
measuring what matters
Institute for Health Metrics and Evaluation



Treatment Pathways in Chronic Disease

Objective: The objective of this study is to characterize the prevalence of different treatment pathways for three chronic diseases: Hypertension, Type II Diabetes, and Depression. We will systematically summarize the treatment pathways observed among patients who have at least 3 years of continuous observation and persistent treatment following initiation. We will stratify the results by year to evaluate temporal trends, and will further stratify by data source to determine if treatment pathways vary by population, geography, and data capture process.

Rationale: While numerous treatment guidelines exist for chronic conditions, there is a paucity of data on the real-world treatment pathways that patients experience in practice. Understanding these pathways is essential for establishing context around questions of drug utilization, effectiveness, and adherence.

Project Leads: Patrick Ryan, Jon Duke, George Hripcsak, Martijn Schuemie, Nigam Shah

Coordinating Institution(s): Janssen R&D, Columbia University, Regenstrief Institute, Stanford University

Additional Participants:

Full Protocol: [Hypertension Treatment Pathways 12-4-2014](#)

Initial Proposal Date: 12/3/2014

Launch Date: 12/5/2014

Study Closure Date: 12/31/2014

Results Submission: [Email](#) or SFTP

Requirements

CDM: V4 or V5

Database Dialect: SQL Server, Postgres, Oracle

Software: SQL as above, R (optional)

Conclusions (for use-case)

- Generalizable to multiple types of data brokerage activities or institutional CDW cores
- Integration (and automation) within the data brokerage pipeline?
- Define and vet an appropriate level of granularity
- Link brokerage metadata to additional scientific outputs across repositories

AMIA Summits 2018 Activity

Activity

1. Work on topics (45 min total)
 - a. Define the topic
 - b. Describe the reproducibility challenges
 - c. List possible metadata needed
 - d. Identify relevant community resources
2. Regroup and report out (25 min)

<http://bit.ly/InformaticsWorkflow>

Takeaways from AMIA 2018 Workshop

- Data Provenance
- Applicability of diverse policies
- Documentation of limitations and uncertainties

Future Work

WG case statement:

<https://docs.google.com/document/d/1wrpxYnldvJHKN21J70esdFhVEabqizjaNXJePAp2SnM/edit?usp=sharing>

- Participate in the working group
- Vet the framework of elements
- Interested in adoption testing?

Questions & Discussion