Open Science Training: Data Management and Sharing Webinar

American Heart Association

Materials adapted, with permission, from the Center for Open Science

American Heart Association.

1

Welcome!

Sara Jarvis, PhD
Research Accountability & Excellence Manager
Sara.Jarvis@heart.org

Award items that fall under my purview:

- Data plans (for some awards)
- Data deposits
- PMCIDs
- Consultations on data sharing
- Pre-registration for clinical trials

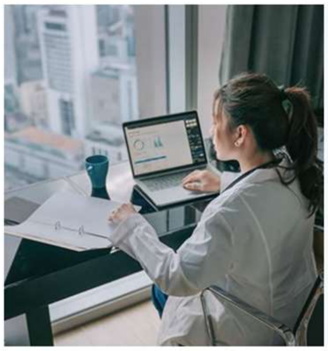
2

American Heart Association.

Award manager

https://professional.heart.org/en/research-programs/awardee-resources

Awardee Resources



Tools to help you activate and manage your award

This page supports AHA awardees and their institutional officers. Access policies related to your award and get help with deliverables and award changes.

Awardees are encouraged to visit the Researchers at Heart alumni site to connect with other awardees, engage with the AHA and explore resources to help you effectively communicate about your project and its outcomes.

Scroll to the bottom of the page…

3

American Heart Association.

Award manager

Contact Us

Policy and Procedural Questions

Email **AHA Awards** (include the award number on the subject line)
or call 214-360-6107, option 2

- May I re-budget funds?
- I wish to replace personnel on my grant.
- May I request a no-cost extension?
- I need to take a leave of absence.
- When is my next progress report or expenditure report due?

Find the AHA Awards Administrator for your state

Login or Technical Help

Email **PC Support**
or call 1-800-875-2562 (toll-free U.S. and Canada)
or 1-703-964-5840 (direct dial international).

- I cannot access my ProposalCentral account.
- I cannot find my award record.
- I get an error message when I upload a report.
- I need to update my institution and contact information, but am unsure how.
- It seems I have more than one login for ProposalCentral. Can I merge them?

4

American Heart Association.

Award Administrators by State

State	Contact Name	State	Contact Name
Alabama	Patricia Otieno	Montana	Maribel Castillo Vasquez
Alaska	Maribel Castillo Vasquez	Nebraska	Maribel Castillo Vasquez
Arizona	Maribel Castillo Vasquez	Nevada	Maribel Castillo Vasquez
Arkansas	Cierra Vaughn-Smith	New Hampshire	Katherine Boren
California	Maribel Castillo Vasquez	New Jersey	Katherine Boren
Colorado	Maribel Castillo Vasquez	New Mexico	Maribel Castillo Vasquez
Connecticut	Katherine Boren	New York	Katherine Boren
Delaware	Katherine Boren	North Carolina	Patricia Otieno
District of Columbia	Patricia Otieno	North Dakota	Maribel Castillo Vasquez
Florida	Patricia Otieno	Ohio	Matilda Mokake
Georgia	Patricia Otieno	Oklahoma	Maribel Castillo Vasquez
Hawaii	Maribel Castillo Vasquez	Oregon	Maribel Castillo Vasquez
Idaho	Maribel Castillo Vasquez	Pennsylvania	Patricia Otieno
Illinois	Cierra Vaughn-Smith	Puerto Rico	Patricia Otieno
Indiana	Cierra Vaughn-Smith	Rhode Island	Katherine Boren
Iowa	Cierra Vaughn-Smith	South Carolina	Patricia Otieno
Kansas	Maribel Castillo Vasquez	South Dakota	Maribel Castillo Vasquez
Kentucky	Matilda Mokake	Tennessee	Matilda Mokake
Louisiana	Cierra Vaughn-Smith	Texas	Cierra Vaughn-Smith
Maine	Katherine Boren	Utah	Maribel Castillo Vasquez
Maryland	Katherine Boren	Vermont	Katherine Boren
Massachusetts	Katherine Boren	Virginia	Patricia Otieno
Michigan	Matilda Mokake	Washington	Maribel Castillo Vasquez
Minnesota	Cierra Vaughn-Smith	West Virginia	Patricia Otieno
Mississippi	Cierra Vaughn-Smith	Wisconsin	Cierra Vaughn-Smith
Missouri	Cierra Vaughn-Smith	Wyoming	Maribel Castillo Vasquez

5

American Heart Association.

Association's Open Science Policies

- Open Data: Any factual data that is needed for independent verification of research results must be made freely and publicly available in an Association-approved repository as soon as possible, and no later than the time of an associated publication or the end of the award period (and any no-cost extension), whichever comes first.

- Public Access: The Association requires that all journal articles resulting from Association funding be made freely available in PubMed Central (PMC) and linked to an Association award within 12 months of publication.

6

Misconceptions about data deposits into a repository:

Misconception #1:

Only "big data", genomic data, or RNA seq should be deposited into a repository.

Answer: All factual data collected during your award period should be deposited. This includes (but not limited to) data such as: anthropometric measurements, blood/hormone analyses, and any physiologic variables.

Misconception #2:

I have already published my data so I do not have anything to deposit.

Answer: A publication by itself, even if the publisher requires supplemental material deposited on their website, does not constitute a data deposit. These data need to be deposited into an Association approved repository.

Misconception #3:

My award has ended but I have not had time to publish yet, so I will wait until it is published then deposit the data.

Answer: When your Data Deposit Confirmation is due, your data deposit is due. The data should be made freely and publicly available in an Association approved repository, per the award agreement.

7

Why is Open Science important?

Advances rigorous and reproducible research

- Enable validation of research results
- Make datasets accessible
- Accelerate research directions
- Increase opportunities for reuse, citation, and collaboration

Promotes public trust in research

- Foster transparency and accountability
- Demonstrate stewardship over dollars spent
- Maximize research participants' contribution

8

Updated October 23, 2024
For awards beginning August 1, 2023 and later

Association's Open Science Policy Compliance

Data Plan

Submit with application

Update as needed

Preregistration

Required for funded clinical trials

Encouraged for all studies making inferential claims

Preprints

Encouraged, but not required

Allows for early communication of findings

Submit Publications to PubMed Central

Journal articles must be made freely available and linked to an Association award within 12 months of publication

Deposit Data into an Association-Approved Repository

Due as soon as possible, and no later than the time of an associated publication or the end of the award period (or NCE)

Any factual data needed for independent verification of research results

Data Deposit Confirmation

Due as soon as possible, and no later than the time of an associated publication or the end of the award period (or NCE)

Available in Deliverables section of ProposalCentral

9

American Heart Association.

Association's Data Plan elements

- Describe the factual data necessary to independently verify your research findings.
  - This should just be a few sentences to describe your data and the accompanying metadata.
  - Do not reference potential publications or presentations.
- List the approximate date the data will be shared.
- Describe any limits to data sharing that might be required.
- Outline where the data be shared. That is, which Association-approved repository(ies).

10

In this webinar, we will learn how to:

- ✓ Identify data management requirements
- ✓ Create a data management plan
- ✓ Create a file structure convention
- ✓ Create a file naming convention
- ✓ Create data dictionary or codebook
- ✓ Create a project README file

11

Terms we will use:

- **Data management plan**: document that describes how you will manage, describe, analyze, and store research data
- **Data dictionary / codebook:** document describing elements of a dataset
- **README file:** document describing a research project including the context of your project, the materials involved, how they relate to one another, naming conventions, sources

12

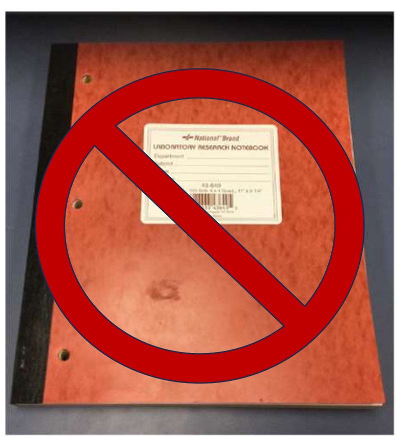
Data management

American Heart Association.

13

American Heart Association.

Data management plan



14

Data management plan: plan a home for your data in advance

Use a data repository from the Association-approved list

- Persistent identifiers for your data (like DOI)
  - Unique and citable
  - Prevents "link rot"
- Persistent access
- Preservation
- Backup
- Management of access
- Versioning
- Licensing

15

Create an informative directory structure for your files

- Directories (main folder) and subdirectories (nested folders) organized to make research materials discoverable and understandable
- Create subdirectories for like materials: separate data, code, and results
- Locations should be distinctive, consistent, and informative:
  a. What it is
  b. Why it exists
  c. How it relates to other files

16

American Heart Association.

Directory structure and file naming

```
.
|-- CITATION
|-- README
|-- LICENSE
|-- requirements.txt
|-- data
|     -- birds_count_table.csv
|-- doc
|     -- notebook.md
|     -- manuscript.md
|     -- changelog.txt
|-- results
|     -- summarized_results.csv
|-- src
|     -- sightings_analysis.py
|     -- runall.py
```

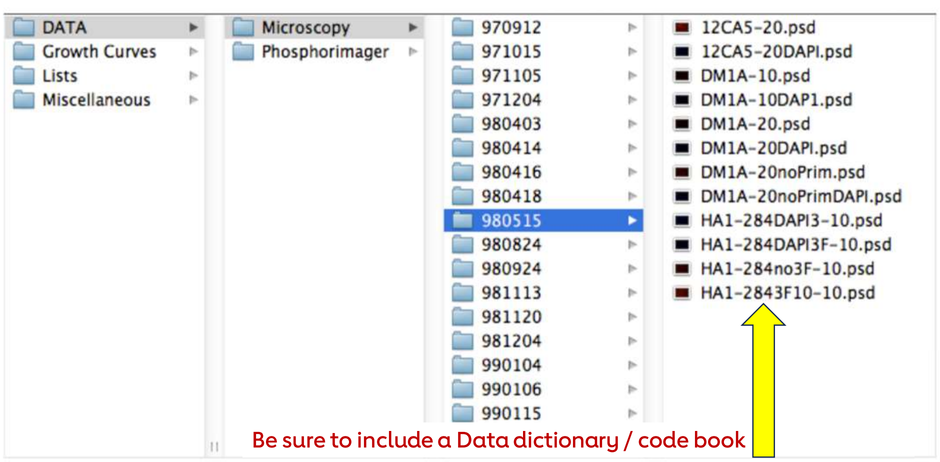
Source: kbroman.org/dataorg/

17

American Heart Association.

Source: https://guides.library.stanford.edu/data-best-practices/name-files

Directory structure and file naming



Be sure to include a Data dictionary / code book

18

Source: https://guides.library.stanford.edu/data-best-practices/name-files

Directory structure and file naming

Example

The example photo shown on the right above was named using this convention as

FR3S.140623.129C.2653.W.JPG

How does this translate?

- FR3 = study site FR3
- S = shallow
- 140623 = June 23, 2014
- 129 = tile number 129
- C = covered treatment
- 2653 = photo number assigned by camera
- W = whole tile

19

Data sharing

20

What does it look like to share your research?

There is a spectrum of sharing research, including:

- Raw data, code, and materials
- Selected data, code, and materials
- Brokered, embargoed, or other limited sharing
- Rich metadata only

21

Research sharing occurs on a spectrum:

“As open as possible, as closed as necessary”

Sharing your research does not mean

- Giving away proprietary information
- Violating privacy of study participants; coercing participants
- Losing control of how your research is reused
- Sharing everything

22

American Heart Association.

Research sharing 101

What to share?

- Data and code necessary to verify findings and reproduce results
- Data and code that might be valuable to other researchers/policy makers
- Data and code which cannot be (easily) regenerated

Why share?

- Funder or publisher mandates
- Citation benefits (ex: Piwowar 2013, https://doi.org/10.7717/peerj.175)
- Preserve long-term access to data

How to share?

- Choose open, persistent, and non-proprietary file formats
- Create and share documentation to enable reuse
- Include data citations of source data
- Create rich metadata

23

American Heart Association.

There are many benefits to sharing

Sharing your research can:

- Prevent loss of access to your data and materials
- Improve the transition when hiring new staff to your research
- Increase the impact and reuse of your research
- Gain credit for your work

24

American Heart Association.

Preserve access to your research:

Repositories:
- Persistent identifiers for your data (like DOI)
  - Unique & citable
  - Prevents "link rot"
- Persistent access
- Preservation
- Backup
- Management of access
- Versioning
- Licensing

Data License:
- Consider Creative Commons licenses for data and text
- Consider either CC-0 + attribution or CC-BY 4.0 [see our FAQs section]
- Guidance on licenses can be found here: https://creativecommons.org/

Code License:
- Consider open source license such as MIT, BSD, or Apache license
- Guidance on software licenses by Karl Broman: http://kbroman.org/steps2rr/pages/licenses.html and Open Source Initiative: https://opensource.org/licenses

25

American Heart Association.

What kind of repositories are there?

Disciplinary repositories
- Targeted disciplines
- Rich discipline metadata

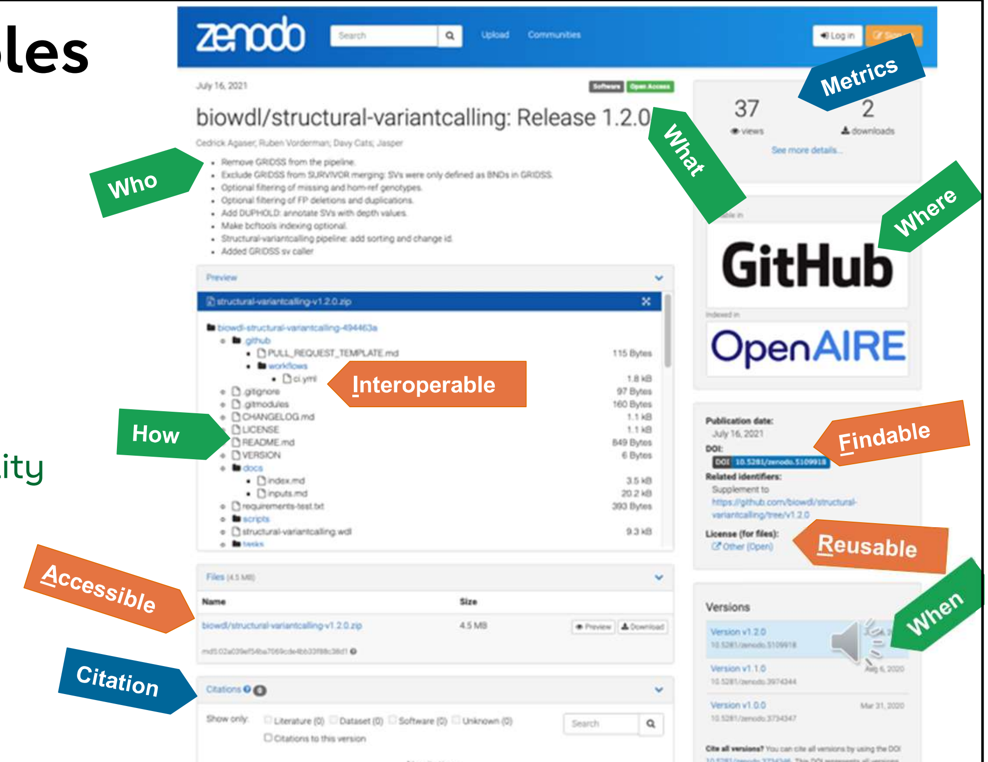
Generalist repositories
- Broad and interdisciplinary
- Commonly used metadata

Institutional repositories
- Interdisciplinary, but limited to an institution
- Commonly used metadata

26

FAIR principles in data

- Metadata quality
  - FAIR
    - Findable
    - Accessible
    - Interoperable
    - Reusable
- Documentation quality
  - Who, What, Where, When, How
- Reuse indicators
  - Metrics
  - Citations



Ideally, upload all your materials in one place

File tree

- Central place to upload and organize files
- Built-in version control
- Previous version remains for review and download
- Editable text files (.txt, .csv)

Tell others how to use your research

License

- Communicates who and how your project is reused
- License picker for common licenses
- CC-BY 4.0 ensures people are prompted to give attribution
- Licenses are controlled at the component level

29

https://kbroman.org/dataorg/pages/dictionary.html

Tell others how to use your research

Data dictionary / codebook

- The exact variable name as in the data file
- A version of the variable name that might be used in data visualizations
- A longer explanation of what the variable means
- The measurement units; today's gold standards may be different in 5 years
- Expected minimum and maximum values

30

Data dictionary / codebook

Part A: Define Data Sets: The table below, "Sample Data", shows sample data of study participant records collected at two different points in this study. The columns shown the headers (e.g., "ID," "Date," etc.).

Part A - Sample Data

ID	Date	Height	Weight	Age	Sex	Plan	Phase	X
CV1536	2023-07-15	155.7	54.5	21	F	A	PRE	100
CV1362	2023-07-18	167.9	63.3	32	M	B	PRE	101
CV1482	2023-07-15	165.0	52.3	45	F	B	PRE	80
CV1603	2023-07-15	178.2	71.9	64	F	A	PRE	43
CV1761	2023-07-07	173.9	85.1	25	M	C	PRE	62
CV1934	2023-07-23	188.2	82.6	37	M	C	PRE	76
CV1536	2023-10-13	155.4	52.0	21	F	A	POST	543
CV1362	2023-10-16	159.3	65.0	33	M	B	POST	724
CV1482	2023-10-13	165.0	52.0	45	F	B	POST	1000
CV1603	2023-10-13	178.0	72.0	64	F	A	POST	928
CV1761	2023-10-05	173.0	84.9	25	M	C	POST	658
CV1934	2023-10-18	189.4	82.6	38	M	C	POST	1167
CV1536	2023-10-07	155.7	55.3	21	F	A	POST	1614
CV1362	2023-10-07	167.9	62.1	33	M	B	POST	1394
CV1482	2023-10-07	165.0	54.0	45	F	B	POST	1765

31

Data dictionary / codebook

Part B: Create the **Data Dictionary**: The Data Dictionary provides detailed information about file contents and serves as a reference guide for the collected data, enabling those responsible for collection to do so in a consistent manner throughout the project. This Data Dictionary is crucial to help others understand your data. Below is an example Data Dictionary that defines the meaning of the variables in the **Sample Data spreadsheet**. The Data Dictionary may also include information about the type of data (i.e., file type or quantitative/qualitative) and size (i.e., file size, text or numerical field length), among any other information deemed important for understanding what a variable means and for understanding the format in which it should be collected.

Data Dictionary (for Part A - Sample Data)

Field Name	Description	Data Type	Expected Length of the Variable	Units	Variables	Comments
ID	Subject Unique ID	text	6	N/A		Two alpha followed by 4 numerical
Date	Date recorded	date	10	N/A		yyyy/mm/dd
Height	Height of Subject	number	4	cm		
Weight	Weight of Subject	number	3	kg		
Age	Age of subject	number	2	yr		
Sex	Biological sex at birth	text	1	N/A	M = Male; F = Female	
Plan	Treatment Plan/Protocol	text	1	N/A	A = Plan A; B = Plan B; C = Plan C	treatment protocols/interventions, as defined in study documentation
Plan A	Medication A	number	2	mg	30-90 mg	Calculated based on body weight
Plan B	Medication B	number	1	mg	8 mg	
Plan C	Medication C	number	2	mg	50 mg	
Phase	Timing of Testing	text	3-4	N/A	PRE = Before treatment/intervention POST = After treatment/intervention	12 weeks between PRE and POST
X	Measurement of physiologic variable X	number	2-4	a.u.		this is a fictitious variable for the sake of this example

Your **Data Dictionary** will be uploaded into the data repository, along with the README file and data files.

Click here: to download an Excel template file to begin creating your Data Dictionary

Your **Data Dictionary** will extract the column titles from Part A Sample Data File to become your Field Name titles. Provide the description, data type, expected length of the variable, the units, what the expected variables would be, and any comments. Additional columns may be added as the researcher sees fit.

This **metadata** provides important context about the shared data, enabling users to search, browse, and filter data sets to locate the data in which they are interested.

32

Where to find this information…

https://professional.heart.org/en/research-programs/awardee-resources/open-science-policy-statements-for-aha-funded-research

Scroll to the bottom of the page…

AHA Open Science Policy Resources

- Frequently Asked Questions
- Steps for Open Science Compliance (PDF)
- How to deposit your AHA publication into PubMed Central (video)
- HIPPA Compliance for Open Data Policy (PDF)
- AHA Approved Data Repositories
- Sample Data Plans
- Instructions for adding/modifying an Open Science Data Plan deliverable
- Data Deposit Workbook (XLSX)
- How to add a Data Deposit Confirmation deliverable

The **Data Deposit Workbook** contains:

- Links to videos to define terms
- Links to other resources like licensing, the Association's approved repositories, and other helpful information
- Downloadable templates for a Data Dictionary and a README file, files needed to accompany your data deposit
- Link to instructions for how to complete a Data Deposit Confirmation in ProposalCentral

Tells others about your research and how to use your research:

README file (metadata about your data)

This file has four main components:

1. General information
2. Data and file overview
3. Sharing and access information
4. Methodological information (brief)

American Heart Association.

Association's README file template

README file

A README file is often the first file someone will open in your data repository. The purpose of this file is to provide a roadmap for your data deposit and contextualize your project by providing metadata for the inquirer.

Instructions: Fill out this document to its entirety and upload with your data deposit in an American Heart Association approved repository.

Each data deposit into a repository requires a README file and a Data Dictionary.

General information

Date of data deposit submission

1. PI's name: Enter PI's name
2. PI's ORCiD ID: Enter PI's ORCiD ID
3. Association award number: Enter Association award number
4. Association award doi: Enter Association award doi
5. Association award title: Enter Association award title
6. Date range of data collection: Enter data collection date range
7. List of up to six key words to describe data topic:
   a. Enter Keyword 1
   b. Enter Keyword 2
   c. Enter Keyword 3
   d. Enter Keyword 4
   e. Enter Keyword 5
   f. Enter Keyword 6
8. Other: Enter other text, or N/A

Data and file overview

1. For each file name, a short description of what data it contains. Enter list of file names, descriptions, what files contain, and dates that files were created
2. Other: Enter other text, or N/A

Note the data files you upload should be in a common format, accessible by anyone (Excel, Word, pdf, CSV, txt). Data files should not require a specialized software or application to open. Any analyses performed should be described below in sharing and access information (software, hardware with version).

35

35

American Heart Association.

Association's README file template

The Association's Open Science Policy states: Any factual data that is needed for independent verification of research results must be made freely and publicly available in an Association-approved repository as soon as possible, and no later than the time of an associated publication or the end of the award period (and any no-cost extension), whichever comes first.

Sharing and access information

1. Licenses on data (The Association recommends CC-0 + attribution or CC-BY 4.0): Enter data license
2. Licenses on source code: Enter software or code license, or N/A
3. Links to other data deposits/repository links: Enter links to other data repository(ies), or N/A
4. Recommended citation for the data: Enter citation information, or N/A

Methodological information

1. Description of model:
   a. Human, animal: Enter human or animal
      i. Animal model, enter additional information: If animal, describe, or N/A
   b. Sex: Enter sex
   c. Sample size: Enter sample size
   d. Age: Enter age, age range
   e. Disease state: Enter disease state/describe population
   f. Interventions: Enter intervention, or N/A
   g. Comparisons: Enter comparisons, or N/A
   h. Other: Enter other text, or N/A
2. Description of methods for data collection (may include links or references to publications or other documents): Enter text for data collection, may include links or references.
3. Description of methods used for processing data (how were data processed from raw data to analyzed data): Enter text for methods used for processing data, may include links or references.
4. Describe any software or instruments needed to interpret or understand the data/results of your study. Please be specific to include software and hardware version numbers. Enter text for software or instruments needed to interpret or understand data.
5. Other: Enter other text, or N/A

36

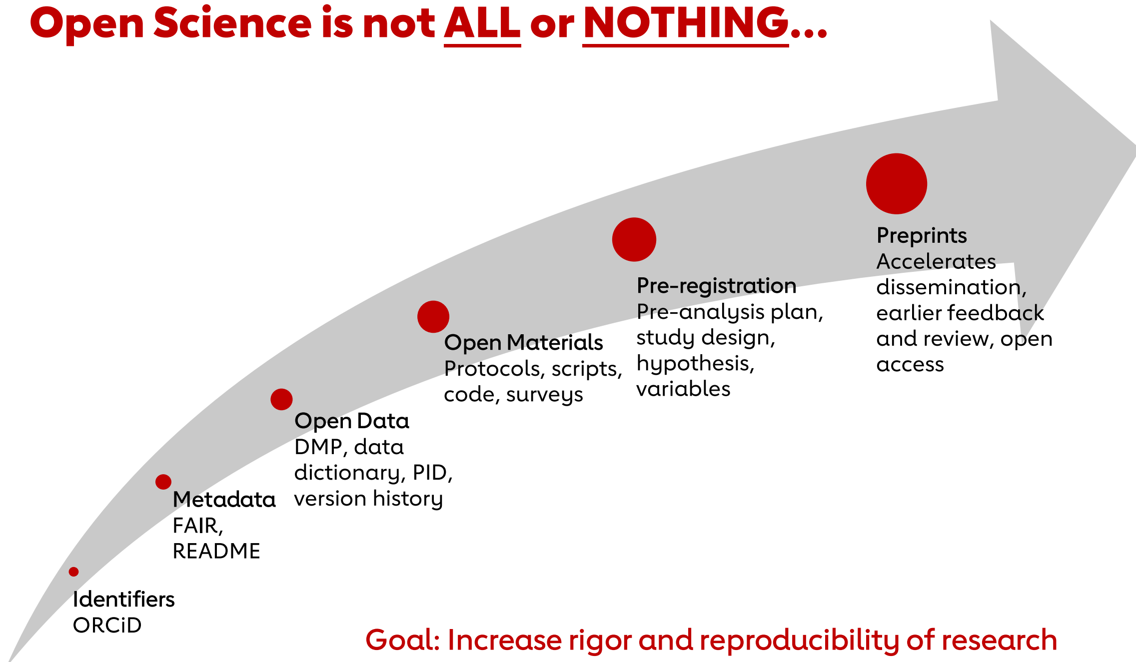
36

Key points

American Heart Association.

37

Open Science is not ALL or NOTHING…



Goal: Increase rigor and reproducibility of research

38

Now what? What to do after today...

- Who to contact if you have questions about:
  - Your award: **Refer to Resources website for your Award Manager**
  - Open Science: **Sara Jarvis, PhD**
- A pdf version of the slides is also available on the website
- The **Data Deposit Workbook** is available for download
  - Your data deposit into an Association approved repository must include both a **Data Dictionary** and a **README file** such that the appropriate metadata is included with your deposit
- Things to think about:
  - Ideas: revise data management plan
  - Consider a folder and file naming convention
  - Do you have specific questions about your data and need a consultation appointment?

39

Thank You.

40