



Data management

CDS 2022 BTS workshop

Rick Gilmore & Kasey Soska

2022-04-21 07:53:21

Acknowledgements

- Support from NICHD, NIH/OD, NIMH, & NIDA via R01HD094830-01; NSF via 2032713; the LEGO Foundation; & the Alfred P. Sloan Foundation
- Karen Adolph, Cathie Tamis-LeMonda, Orit Hertzberg, Tiger Teng

Overview

- A cautionary tale
- (Hyper)active Curation
- Lessons learned

A cautionary tale



(NYU Health Sciences Library, 2013)

(Hyper)active curation

Journal of eScience Librarianship

putting the pieces together: theory and practice

[Home](#) > [Lamar Soutter Library](#) > [Journal of eScience Librarianship](#) > [Vol. 10 \(2021\)](#) > [Iss. 3](#)

(Hyper)active Data Curation: A Video Case Study from Behavioral Science

 Download

[Kasey C. Soska](#), *New York University*
[Melody Xu](#), *New York University*
[Sandy L. Gonzalez](#), *New York University*
[Orit Herzberg](#), *New York University*
[Catherine S. Tamis-LeMonda](#), *New York University*
[Rick O. Gilmore](#), *The Pennsylvania State University*
[Karen E. Adolph](#), *New York University*

Follow
Follow
Follow
Follow
Follow
Follow
Follow

 [Additional files available below](#)

120 DOWNLOADS

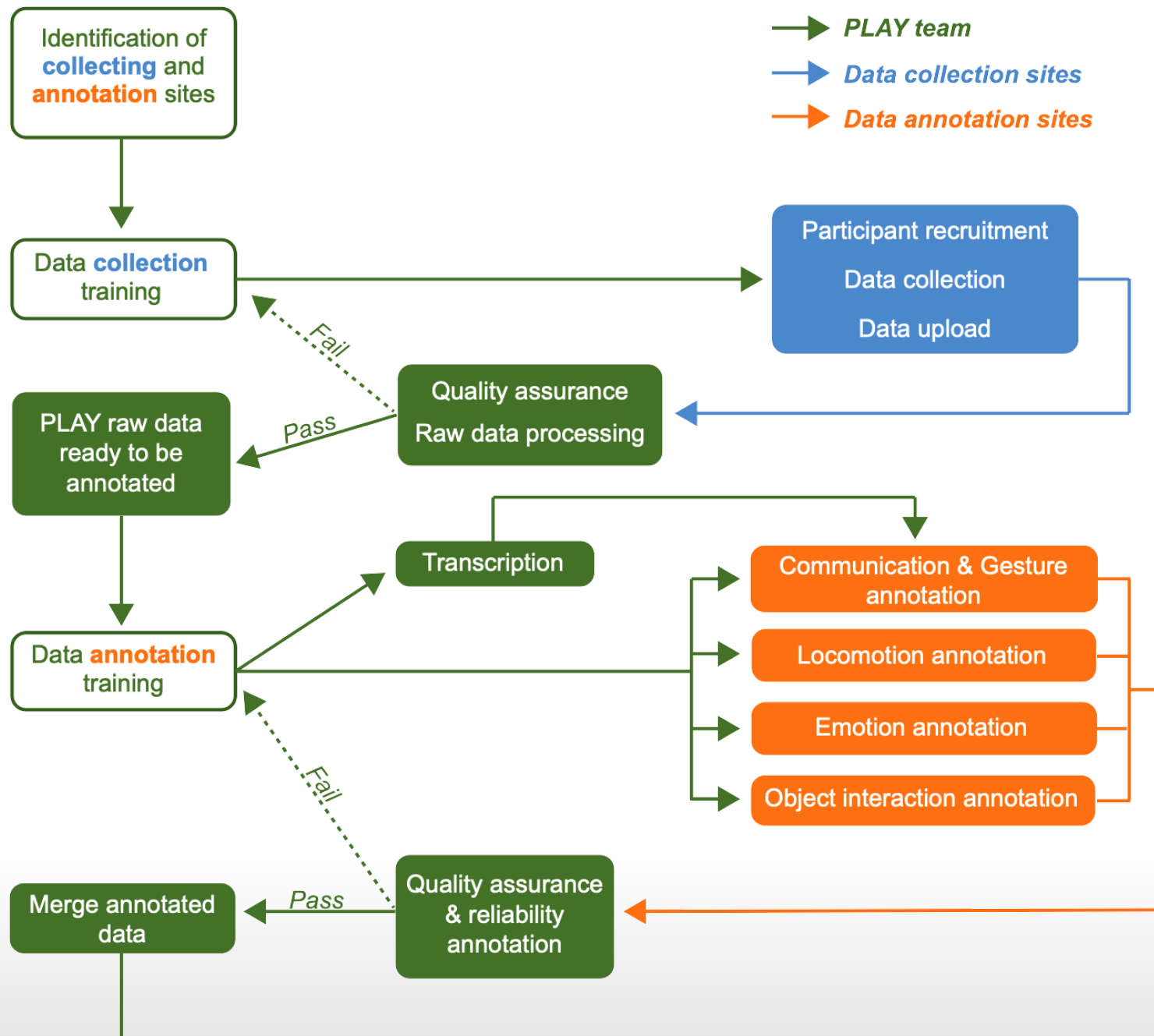
Since August 11, 2021

[\(Soska et al., 2021\)](#)

What is it?

- Embed data *curation* within data *collection* workflow
- Rigorous & thorough quality assurance (QA) during collection
- Curate data with specific sharing target in mind

Workflow



Collection

Coding (video annotation)

Quality assurance (video)

- “Heavy” vs. “light” QA
- Semi-automated QA reports from Databrary
 - https://github.com/PLAY-behaviorome/workflow/tree/master/session_qa_reports
 - Not public

NYU collection volume

2021-02-16 18:17:00

Spreadsheet & Video Checks

Scroll left/right or up/down within tables to view more data.

Spreadsheet data

Name checks












Spreadsheet variable checks

Video checks

[illegible]

Data export & cleaning (surveys)

- <https://github.com/PLAY-behaviorome/KoBoToolbox>
- Series of enumerated R Markdown documents
 - Reproducible
 - Documents gathering, cleaning steps

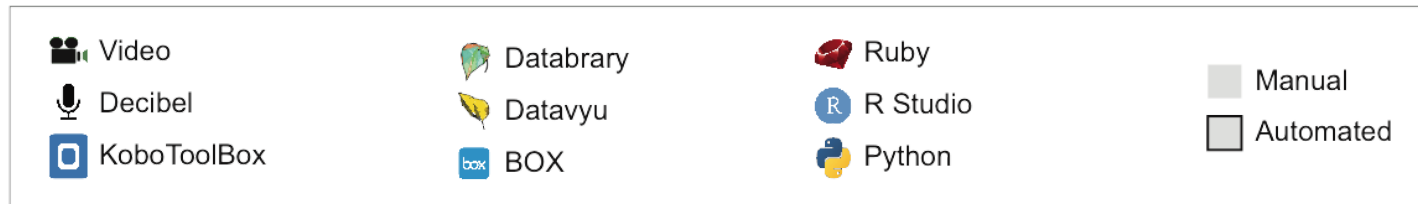
 01-download_from_KoBo.Rmd	Create step-by-step workflow with separate documents.	26 days ago
 01-download_from_KoBo.html	Create step-by-step workflow with separate documents.	26 days ago
 02-save_raw_csvs.Rmd	Create step-by-step workflow with separate documents.	26 days ago
 02-save_raw_csvs.html	Create step-by-step workflow with separate documents.	26 days ago
 03-split_mbcidi_others.Rmd	Create step-by-step workflow with separate documents.	26 days ago
 03-split_mbcidi_others.html	Create step-by-step workflow with separate documents.	26 days ago
 04-remove_identifiers.Rmd	Create step-by-step workflow with separate documents.	26 days ago
 04-remove_identifiers.html	Create step-by-step workflow with separate documents.	26 days ago
 05-conduct_initial_qa.Rmd	Create step-by-step workflow with separate documents.	26 days ago
 06-make_aggregate_csv.Rmd	Create step-by-step workflow with separate documents.	26 days ago
 06-make_aggregate_csv.html	Create step-by-step workflow with separate documents.	26 days ago

<https://github.com/PLAY-behaviorome/KoBoToolbox>

- Final CSV uploaded to Databrary
- ```
play_data <-
databraryapi::read_csv_data_as_df(session_id =
51539, asset_id = 366382)
```

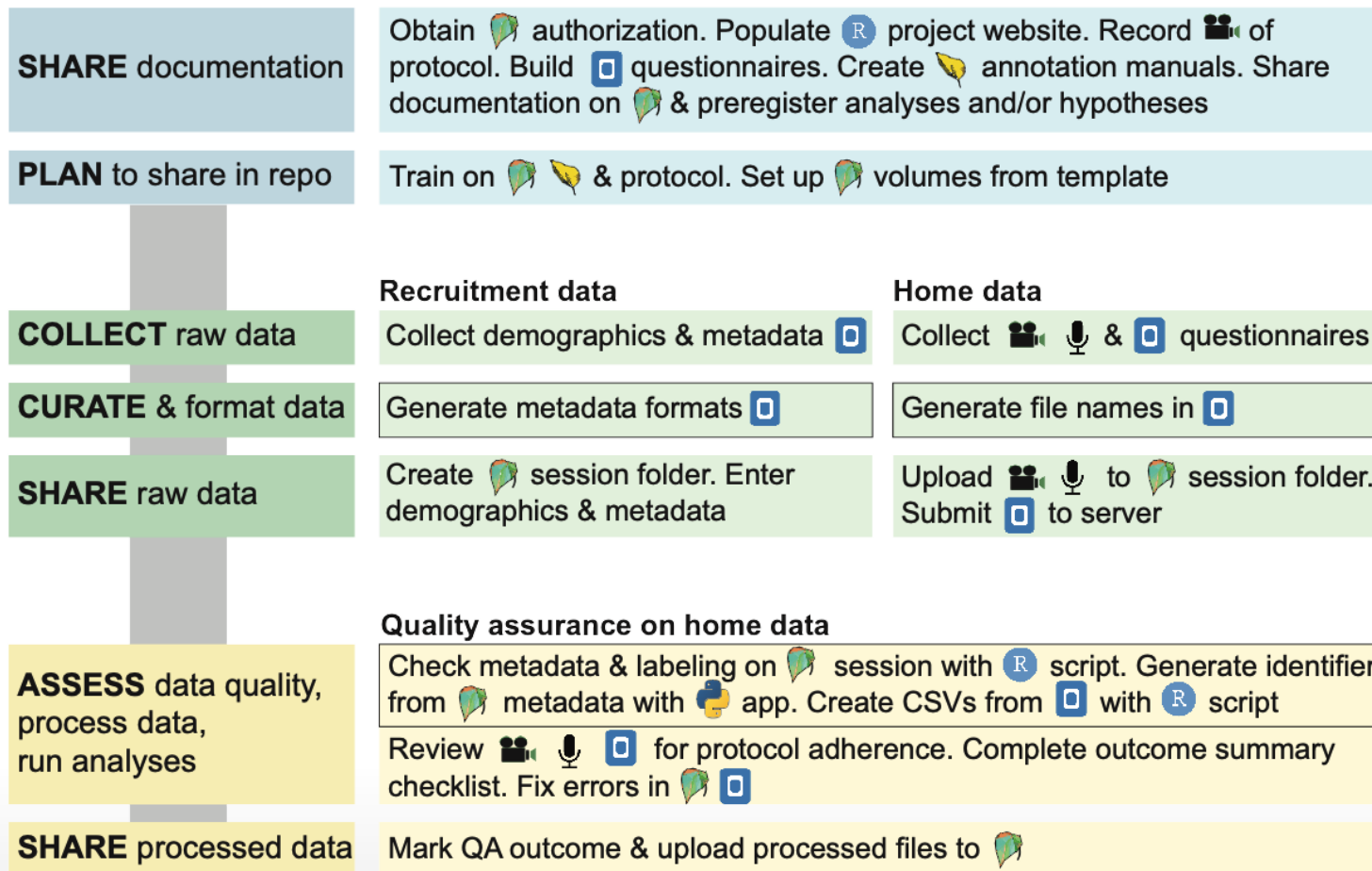


# Summarized



## Hyperactive curation

## PLAY path to share



## Annotation & transcription

## Interobserver reliability

# In-process

- Full protocol on <https://play-project.org>
- Considering migration of protocol to [bookdown](#)
- Individual-level files
  - Including CHAT export of transcripts
- Clean, aggregate MB-CDI data
- From data dictionaries to open schemata

Lessons learned



*"...psychologists tend to treat other peoples' theories like toothbrushes; no self-respecting individual wants to use anyone else's."*

(Mischel, 2009)

*"The toothbrush culture undermines the building of a genuinely cumulative science, encouraging more parallel play and solo game playing, rather than building on each other's directly relevant best work."*

(Mischel, 2009)

“...psychologists tend to treat other peoples’ theories  
data, data management practices, tasks, displays...like  
toothbrushes...”

*Cheeky open and reproducible developmental science  
advocates who want to create a more cumulative science*

# We don't talk about (you know)...



(DisneyMusicVEVO, 2021)



# Plan your work; work your plan

- You have to curate data for yourself, so...
- Clear to you == (often) clear to others
- Curate with specific target in mind

- Automate as much as possible
  - Script, use APIs
  - Exploit the web
  - Consistency is the ~~hobgoblin of little minds~~ key to successful automation!

- Test workflow at every step/look at data
- Expect to iterate
- Don't make the perfect the enemy of the good
- *Meaningful* data management plans increasingly required by funders
- Make data a first class product

# Come PLAY with us!

- Share best practices
- Talk about “you know”
- Let’s solve as-yet-unsolved problems...together
- No reinventing wheels



# PLAY

Play & Learning  
Across a Year

<https://play-project.org>

<https://anhourinthelife.org>

<https://PLAY-behaviorome.github.io/2022-04-21-team-sci-cds>



# Databrary

Discover more, faster

[rog1@psu.edu](mailto:rog1@psu.edu)

<https://databrary.org>  
<https://gilmore-lab.github.io>

# Resources

This talk was produced on 2022-04-21 in [RStudio](https://www.rstudio.com/) using R Markdown and the ioslides framework. The code and materials used to generate the slides may be found at <https://github.com/PLAY-behaviorome/2022-04-21-team-sci-cds/>. Information about the R Session that produced the code is as follows:

```
R version 4.1.2 (2021-11-01)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Monterey 12.3
##
Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
attached base packages:
[1] stats graphics grDevices utils datasets
[6] methods base
##
other attached packages:
```



# References

- DisneyMusicVEVO. (2021, December). We don't talk about bruno (from "encanto"). Youtube. Retrieved from <https://www.youtube.com/watch?v=bvWRMAU6V-c>
- Mischel, W. (2009). Becoming a cumulative science. *APS Observer*, 22(1). Retrieved from <https://www.psychologicalscience.org/observer/becoming-a-cumulative-science>
- NYU Health Sciences Library. (2013, November). Data sharing and management snafu in 3 short acts (higher quality). Youtube. Retrieved from [https://www.youtube.com/watch?v=66oNv\\_DJuPc](https://www.youtube.com/watch?v=66oNv_DJuPc)
- Soska, K. C., Xu, M., Gonzalez, S. L., Herzberg, O., Tamis-LeMonda, C. S., Gilmore, R. O., & Adolph, K. E. (2021). (Hyper)active data curation: A video case study from behavioral science. *Journal of Escience Librarianship*, 10(3). <https://doi.org/10.7191/jeslib.2021.1208>