Working with Other People's Data

Challenges & Considerations

Harvard DataFest 2018
17 - 18 January 2018

About Us

- **Daina Bouquin**
  - Head Librarian, Harvard-Smithsonian Center for Astrophysics
  - daina.bouquin@cfa.harvard.edu
- **Ceilyn Boyd**
  - Research Data Program Manager, Harvard Library
  - ceilyn_boyd@harvard.edu
  - http://hlrdm.library.harvard.edu
- **Barbara Esty**
  - Senior Research Information Specialist, Harvard Business School
  - baesty@hbs.edu

Agenda

- Overview of Considerations & Challenges
- Case Studies
  - Astrophysics: Interoperability and Integrating Multiple Sources
  - Working with Faculty & Their Co-Authors
  - Project Workflows for Humanists & Social Scientists
- Summary, Resources & Discussion

Considerations & Challenges: Acquisition, Use, Reuse & Sharing of Other People's Data

Legal & Ethical

- E.g. Data licensing, terms of use. attribution, sensitivity in using online data

Technical & Usability

- E.g. File format conversion, data completeness, storage and computing resources

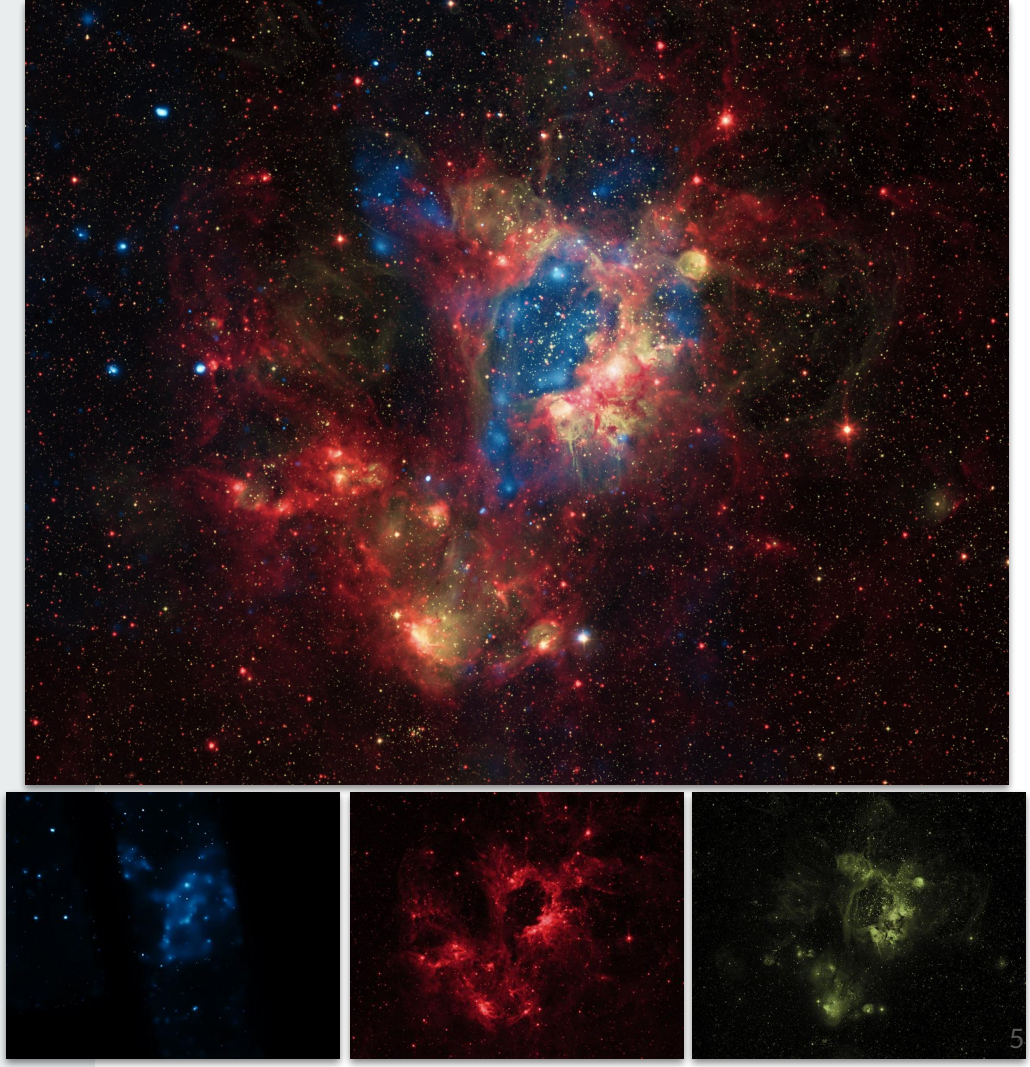
Sustainability

- E.g. Cost of support, bit rot, DOIs

Daina Bouquin
daina.bouquin@cfa.harvard.edu

Case Study: Astrophysics

Interoperability and Integrating Multiple Sources



Superbubble in the
Large Magellanic Cloud
X-ray: NASA/CXC/U.Mich./S.Oey,
IR: NASA/JPL,
Optical: ESO/WFI/2.2-m

Daina Bouquin
daina.bouquin@cfa.harvard.edu

Considerations

Legal & Ethical

- Trustworthiness of each data source
  - Transparency
- Ascribing scholarly credit to all sources
  - Reconciling attribution issues

Technical & Usability

- “Interconnection is not merely linking database queries, but facilitating science and knowledge discovery within multiple data types and formats that correspond to multiple wavelengths and features of astronomical phenomena and **to varying conditions of the instruments that capture them.**” [Wynholds et al.]
- Formats, documentation, software (!)

Sustainability

- Funding and curation environments for missions of varying size
- Versioning support

Barbara Esty
baesty@hbs.edu

Case Study: Business

Working with Faculty and their Co-Authors

When your data has a people problem...

ARTICLE | HARVARD BUSINESS REVIEW | JANUARY–FEBRUARY 2018

Ads That Don't Overstep: How to Make Sure You Don't Take Personalization Too Far

Leslie John, Tami Kim and Kate Barasz

ARTICLE | JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY

Humblebragging: A Distinct—and Ineffective—Self-Presentation Strategy

Ovul Sezer, **Francesca Gino** and **Michael I. Norton**

ARTICLE | HARVARD BUSINESS REVIEW | JANUARY–FEBRUARY 2018

Inclusive Growth: Profitable Strategies for Tackling Poverty and Inequality

Robert S. Kaplan, **George Serafeim** and Eduardo Tugendhat

ARTICLE | CLINICAL PHARMACOLOGY & THERAPEUTICS | JANUARY 2018

Innovation Incentives and Biomarkers

Ariel Dora Stern, Brian M. Alexander and **Amitabh Chandra**

ARTICLE | HARVARD BUSINESS REVIEW | JANUARY–FEBRUARY 2018

More than a Paycheck: How to Create Good Blue-Collar Jobs in the Knowledge Economy

Dennis Campbell, John Case and Bill Fotsch

ARTICLE | JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY

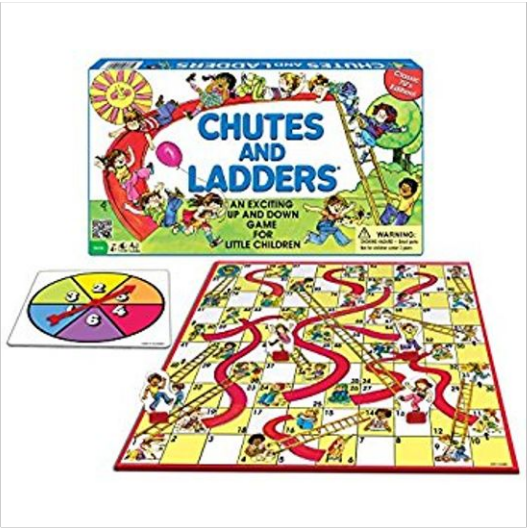
Olfactory Cues from Romantic Partners and Strangers Moderate Women's Responses to Stress

Marlise Hofer, Hanne Collins, **Ashley V. Whillans** and Frances Chen

The Ups and Downs of Working with Co-authors

Ladders

- Subject matter expertise
- Technical skills
- Time
- Resources
- New perspectives



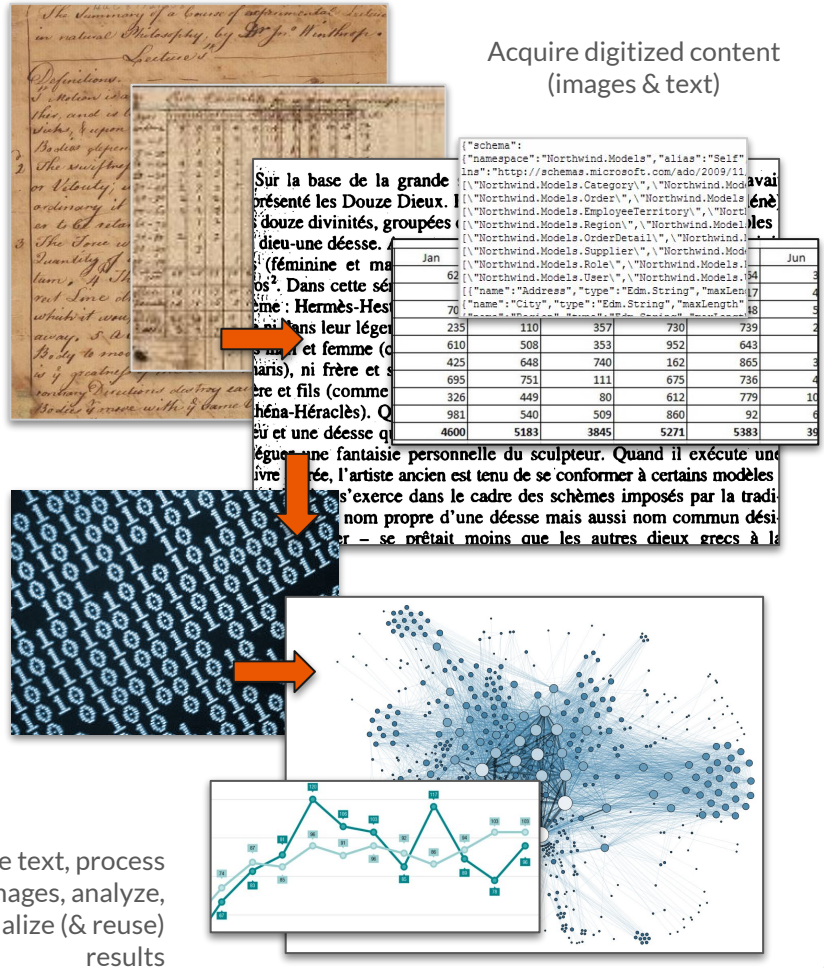
Chutes

- Who's in charge?
- License issues
- Access to data, storage, network…
- Conflicting opinions
- Does this project have an end, owner?

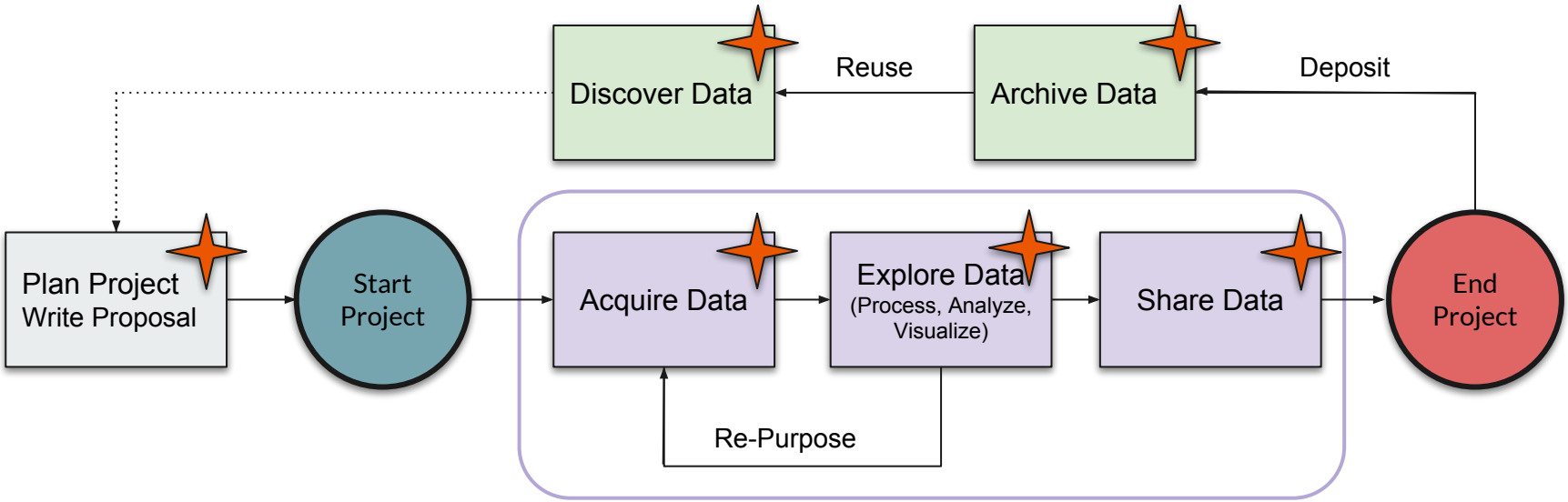
Ceilyn Boyd
ceilyn_boyd@harvard.edu

Workflows for Humanists & Social Scientists

Getting Granular with Text (and Image) Data

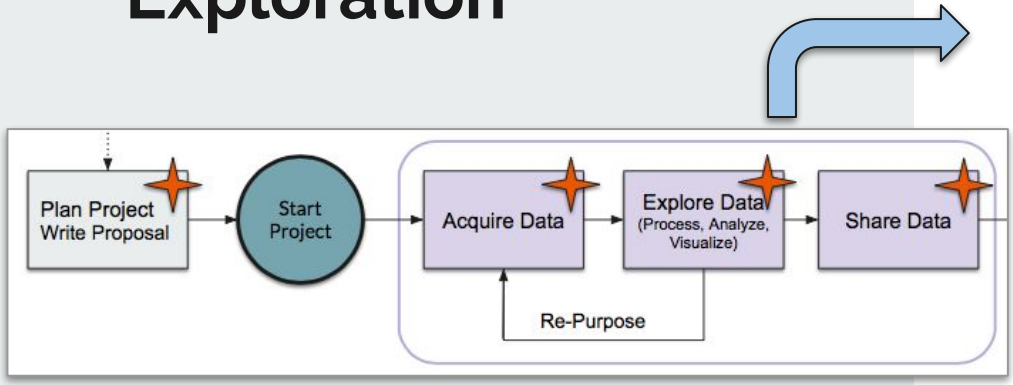


Project & Data Lifecycle Workflow Considerations

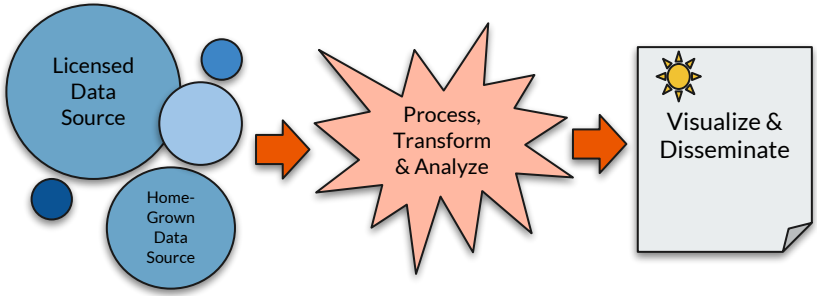


Source: University of Virginia: http://data.library.virginia.edu/data-management/lifecycle/

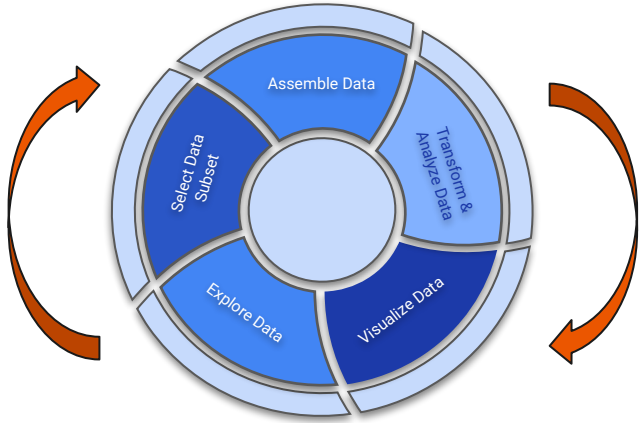
Workflows for Data Storytelling & Exploration



Data Storytelling



Iterative Data Exploration



Workflow Considerations & Challenges

Workflow Component	Legal & Ethical	Technical & Usability	Sustainability
Data sources & Data sinks	• Terms of use • Frequency of use • Copyright restrictions • Limits on public access or sharing with partners	• File formats • Metadata • Limits and scope of data • OCR text quality • Integration with other data sources (e.g. surveys)	• Initial and ongoing costs • Images, data, metadata, or all? • Storage/hosting • Frequency of OCR updates • Data quality and scope
Data transformers & Data analyzers	• Re-identification of data • Terms of use	• File format conversion • Access to knowledge and skills • Walled garden or Open Source • System integration • Interoperability	• Computing resources • Software licensing fees • Walled garden or Open Source? • Reproducibility
Data visualizers & Data dissemination	• Restrictions on quantity of content (e.g. < 100 words of text under copyright)	• Software licensing fees • Access to knowledge and skills	• Software licensing fees • Walled garden or Open Source? • Reproducibility

Summary, Resources & Discussion

Summary

Legal & Ethical	Technical & Usability	Sustainability

• Restrictions on:
  ○ Use
  ○ Transformation
  ○ Dissemination
  ○ Sharing
  ○ Reuse of results
• Copyright
• Attribution & citation
• Data provenance and trustworthiness
	• API limits
	• Dataset size & completeness
	• Data formats, complexity, interoperability & dependencies (e.g. software)
	• Walled-garden problem
	• Resources: Skills, knowledge & technology
	• Data environment
	• Data documentation
		• Time limit for use
		• Vendor support
		• Data update frequency
		• Storage and processing costs
		• Data documentation
		• Funding

Resources

- Data Sources
  - Harvard Dataverse
    - https://dataverse.harvard.edu
  - Zenodo
    - https://zenodo.org/
  - Re3Data (data repository registry)
    - https://www.re3data.org/
  - Harvard Library Social Sciences Data
    - https://guides.library.harvard.edu/datafest2018
  - Numerical Data Collections (Lamont Library)
    - http://hcl.harvard.edu/libraries/lamont/collections/numericdata/
  - Numerical Data Services Dataverse
    - https://dataverse.harvard.edu/dataverse/nds
    - Cross National Time Series (Banks data, ITERATE database, selected survey data)

Resources

- APIs
  - New York Times
    - https://developer.nytimes.com
  - Factiva/Dow Jones
    - https://developer.dowjones.com/site/global/develop/introduction/index.gsp
  - Google Analytics
    - https://developers.google.com/analytics/
  - Google Geocoding
    - https://developers.google.com/maps/documentation/geocoding/start?csw=1
  - Twitter
    - https://developer.twitter.com/
  - Facebook
    - https://developers.facebook.com/
  - Census
    - https://www.census.gov/data/developers/data-sets.html
  - Weather Underground
    - https://www.wunderground.com/weather/api/?ref=twc
  - NASA Open
    - https://api.nasa.gov/

Resources

Tutorials (APIs)

- Using APIs Without Programming
  - Simple ways you can reverse engineer an organization's data to figure out URLs where data is held.
- CodeAcademy's jQuery and AJAX course
  - Good introduction to jQuery and AJAX which can be used to interact with APIs.
- Creating an API-Centric Web App
  - Guide detailing how to create an app that centers around API calls.
- REST API Tutorial:
  - Tutorial on the basics of using RESTful web architecture principles.

Citation

- DataCite - https://www.datacite.org/cite-your-data.html
- Software Citation Example - GitHub/Zenodo integration - https://guides.github.com/activities/citable-code/

Discussion

Thank you.

Contact Us

- **Daina Bouquin**
  - Head Librarian, Harvard-Smithsonian Center for Astrophysics
  - daina.bouquin@cfa.harvard.edu
- **Ceilyn Boyd**
  - Research Data Program Manager, Harvard Library
  - ceilyn_boyd@harvard.edu
  - http://hlrdm.library.harvard.edu
- **Barbara Esty**
  - Senior Research Information Specialist, Harvard Business School
  - baesty@hbs.edu