Research Data Management

Andrew Creamer, Scientific Data Management Specialist,
Brown University Library
Center for Digital Scholarship

Sam Simas, Research Data Management Intern
Graduate School of Library and Information Studies,
University of Rhode Island

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash

Dis-clam-ah! (disclaimer)

Due to the diverse subfields that comprise anthropology, this presentation is not meant to be prescriptive. Each research project has its own idiosyncrasies, and, therefore, it is difficult to generalize a single approach to managing or sharing the research data that you will collect or produce. In all cases, the expectations are that you will collect, create, manage, retain, store, archive, publish, and share any data consistent with applicable laws in the U.S., other jurisdictions of your project, IRB and institutional policies, and within the ethical norms of your discipline and protect culturally sensitive information, and the privacy and confidentiality and safety of your participants.

Contents

- Overview of RDM & RDM Basics
- Case Study
- Possible Outputs
- Post-fieldwork Considerations

Objective: staying organized, saving time

- Use consistent, descriptive file names
  - manhattan-project-oppenheimer-field-notes-20170220-v1.docx
- Imagine what your outputs might be and where you might store them
- Understand copyright/left and common terms of use
- Use adequate metadata to describe your work
- Backup your data: 3-2-1 Rule (multiple copies, multiple formats, different locations; local original copy, local copy, remote copy)
- Keep track of different versions of files (versioning)
- Be aware of data containing personal identifiers and security

What does data mean to you?

Geographic

Maps, GIS, coordinates, locations, landmarks, etc.

Visual

Photographs, postcards, artwork, PNG, JPEG, GIF, glossy photos, print & digital photos, etc.

Audio

CDs, MP3s, WAV, recordings, broadcasts, songs, compositions, scores, etc.

Statistical

Datasets, surveys, census info., demographic, etc.

Video

Film, DVDs, YouTube, Vine, reel-to-reel, 8mm, etc.

Human Subjects

Demographics, ethnography, interviews, journals, etc.

Textual & Hypertextual

Citations, documentation of analog materials, digital scans, OCR text files, documents, spreadsheets, ephemera, publications, appendices, indices, archive of web pages, e.g., html

Why Manage Your Data?

Collecting data for your dissertation

Prioritize documentation and organization for efficiency in discovery, analysis, and writing up findings

Understand your documentation and find what you need easily, even after several months or many years have passed as well as be able access the files when technologies change (e.g., floppy discs, CD-ROMs, DVDs, thumb drives, etc.)

Public and private grant funders increasingly require a data management or public access plan for research materials collected under an award

Increasing number of publishers of monographs and journals requiring making an archive of materials available or detailed indices for peer reviewers and/or scrutiny of published arguments/findings by scholarly communities

Questions for appraising

- What kinds of data do I have?
- How many copies of my files do I have/need?
- What format does my data take?
- What might I want to do with my data in the future?

Created by Gregor Cresnar from Noun Project

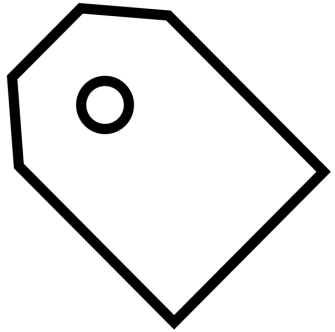
- How will I keep it secure?
- How will I share it? Who will be able to access it?
- What are my options for storing and preserving data?
- How will I organize my data?

Metadata

- Enables data and research to be reused and interpreted for future projects
- Metadata for discovery: location of archive, name of collection, box number/identifier, folder number, file number, document level details
- Rights information
- Information for citation and attribution

Metadata

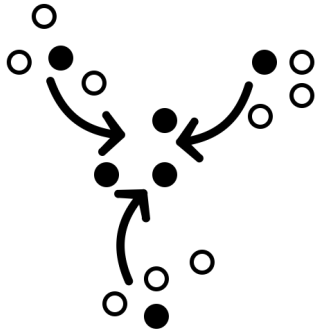
contextual details about your data



Created by Joshua Stearns from Noun Project

Collecting Data

- Spreadsheet management
  - Standardize column content headers (variables) and use of cell content (value)validations
  - Example standardized date: YYYYMMDD
  - Library of Congress Subject Headings (LCSH) for a controlled vocabulary



Created by Evangeline White from Noun Project

- Data dictionary and data glossary to define your variables and values
- Archive/Source information

Professor Schiller's data dictionary explains the variables and values in her spreadsheet containing data collected from state legislature archives. Several states used different terms, so her data dictionary clarifies these variables for any user. This data dictionary is available on her website for her book.

	A	B	
1	Count	Senate Ballot File Variable Name	Senate Ballot Variable Description
2	1	YEAR	The year in which the state legislature voted on the Senate election
3	2	CHAMBER	State legislatures were bicameral with two chambers; House stands for the state House of Repr
4	3	SEAT	Senate terms are six years in length and for this time period started on March 4th and ended six
5	4	ELECTTYPE	Senate elections for expiring terms are lableled general (g) and elections held to fill unexpected
6	5	SEPARATE/JOINT	joint indicates the vote was taken joint session. Voters in joint session were typically but not alwa
7	6	JOINTBALLOT#	For this data, ballot numbering begins with the first joint session ballot labeled 1.
8	7	SENCANDIDATE	Name of candidate for U.S. Senate that state legislator cast a ballot for; NV, absent, not voting,
9	8	ST LEGISLATURE VOTER	Member of state legislature recorded in state or house journal as voting or absent.
10	9	PARTY	Populist
11	10	DISTRICT	Specific district of state legislator; state legislatures apportioned differently across states either b
12	11	COUNTY	Specific district of state legislator; state legislatures apportioned differently across states either b
13			
14			
15			
16		State Completed	Notes
17		California	no county data
18		Delaware	no district data

Choosing Appropriate Formats

Audio

MP3, Broadcast WAVE (with embedded metadata). Uncompressed, final production releases. DSD, PCM 176.4khz, 192khz up to 384kh.

Images

TIFF, JPEG2000, PNG, JPEG/JFIF, DNG, BMP, GIF.

Video

Final production version. Original production resolution and frame rate (i.e. 1080p24; 720p60, etc.).

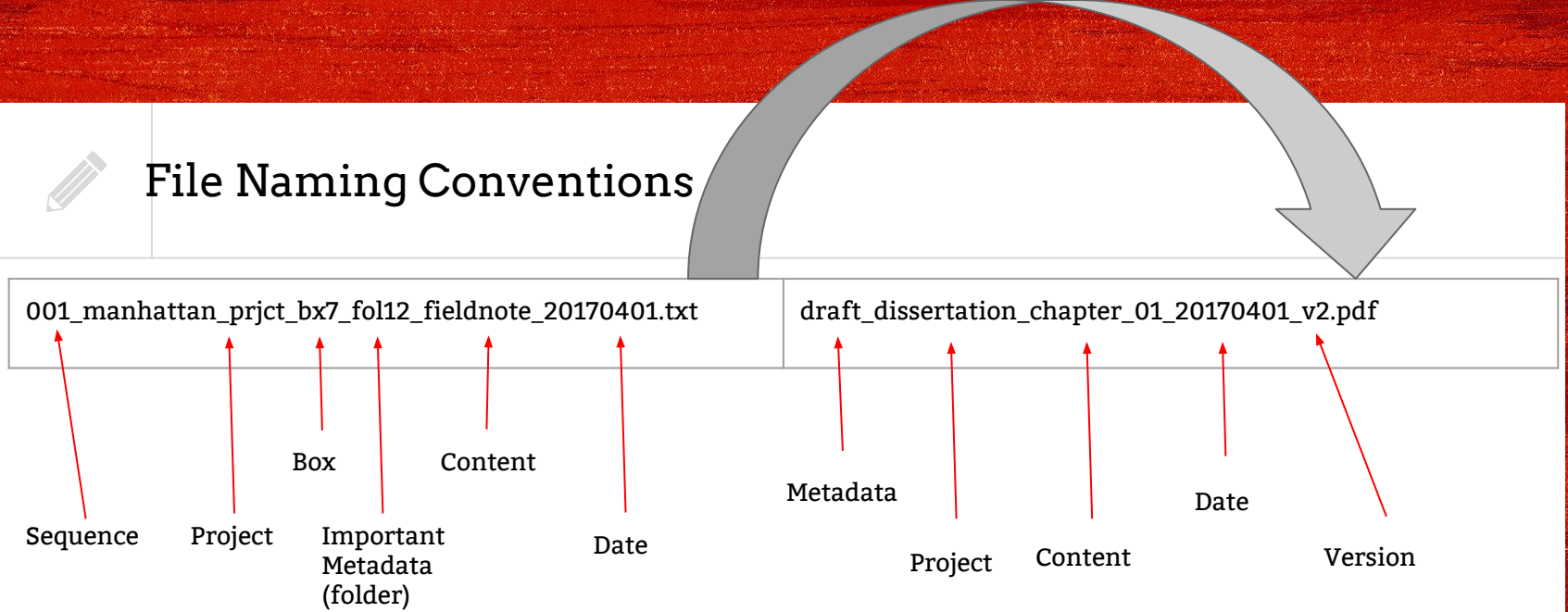
Spreadsheet

CSV

Text

HTML or RTF

File Naming Conventions



Using Folders

- Using folder and subfolder hierarchies
  - Each folder has a README top level file in directory that explains what is in the folder and how it relates to the larger project -- “metadata is a love note to the future”

Files within folders should be numbered 001-010-100-1000

Use tags for files and folders to aid in searching

Storing Data

- 3-2-1 rule: original local copy/local copy stored on a separate drive or network/remote copy of data (Google Drive, DropBox) “Lots of Copies Keep Stuff Safe” – theft, disaster (fire, flood, earthquake, hurricane), crashed computer
- Preservation: migrate data to updated file formats as they evolve or to fresh media for preservation

- Avoid storing primary data on media that easily degrades (optical discs) or is easily lost (thumb drives)
- If you’re sent an optical disk with files from an archive, get the data files off onto a drive, make multiple copies, and store these copies in more than one location (i.e., all your copies on the same drive is not a back up)

Protecting Your Data

- **Backup data** using software or external hard-drives or commercial service
- **Versioning**: use systems like Git that use a timestamp to allow you to return to a previous version or keep a spreadsheet with file name and the changes made, when, and why and a version number (1.1, 1.2…2.0)-make final documents read-only
- Consider depositing your files in the **Brown Digital Repository** (BDR)

Securing Data

- Concerns with taking devices or sensitive data abroad? Use Brown's Information Security Group (ISG) for hardware and software consultations
- Encryption software for PC and MAC is offered by Brown
- Password protected files and folders

Created by Aleksandr Vector from Noun Project

Case Study: So You Want to Make a Map


What does dirty geographic data look like?
United States
Mexico
Guatemala
Nicaragua
Caribbean Sea
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Argentina
Uruguay
North Atlantic Ocean
South Atlantic Ocean
Ireland
United Kingdom
Germany
Poland
France
Austria
Italy
Greece
Tunisia
Algeria
Libya
Mali
Niger
Chad
Burkina Faso
Guinea
Nigeria
Ghana
Gulf of Guinea
Gabon
Angola
Namibia

Barcelona
Barriada Los Galanes, Malaga
Cadiz
Cádiz
Comisaria de Policia de Atarazanas
Cordoba
Córdoba
Distrito Noveno Barcelona
Dos Hermanas
El Palo (Malaga)
estación de M.Z.A
Horta. Barcelona
Hospitalet de Llobregat
Málaga. Paseo de la Rosaleda
Palacio de las Misiones, Barcelona
Piscinas Barcelona
Playa la Barceloneta Barcelona
Plaza Cataluña Barcelona
Pueblonuevo-Peñarroya (Córdoba)
Puesto de Bonares, Huelva
Sevilla
Sevilla Cine Candelaria
Sevilla. Barriada de la Candelaria
Sevilla. La Campana
Sevilla. Plaza del Museo

Arrests-Locations-20170123.xlsx

Find in table

1–200 of 400

	Localidad arresto
1	Barriada Los Galanes, Malaga
2	Distrito Noveno Barcelona
3	Comisaria de Policia de Atarazanas
4	San Roque
5	San Roque
6	San Roque
7	Torremolinos
8	Malaga
9	Torremolinos
10	Nerja
11	Torremolinos
12	Torremolinos
13	Benamocarra
14	Malaga
15	Malaga
16	Marbella

Untitled map

1 view

All changes saved in Drive

Add layer Share Preview

Arrests-Locations-20170123.xlsx

Uniform style

3 rows couldn't be shown on the map. Fix errors highlighted red in the data table. Open data table Dismiss

All items (400)

Base map

Dirty Geographic Data

Dirty Data

is not uniformly organized.

Cadiz  **Cádiz**

El Palo (Malaga)  **Málaga. Paseo de la Rosaleda**

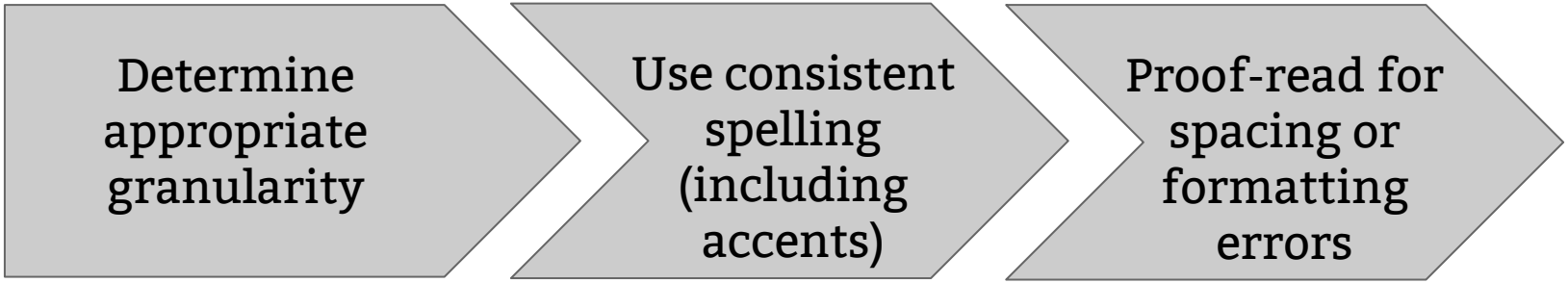
Sevilla  **Sevilla. La campana**

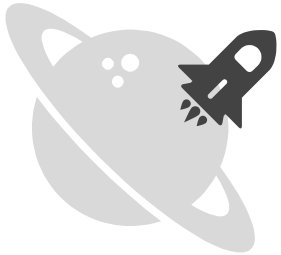
Common Errors

- **Accents**
- **Formatting**
- **Highlighting**
- **Spacing**
- **Spelling**

GRANULARITY

How to fix ‘em


Determine appropriate granularity
Use consistent spelling (including accents)
Proof-read for spacing or formatting errors

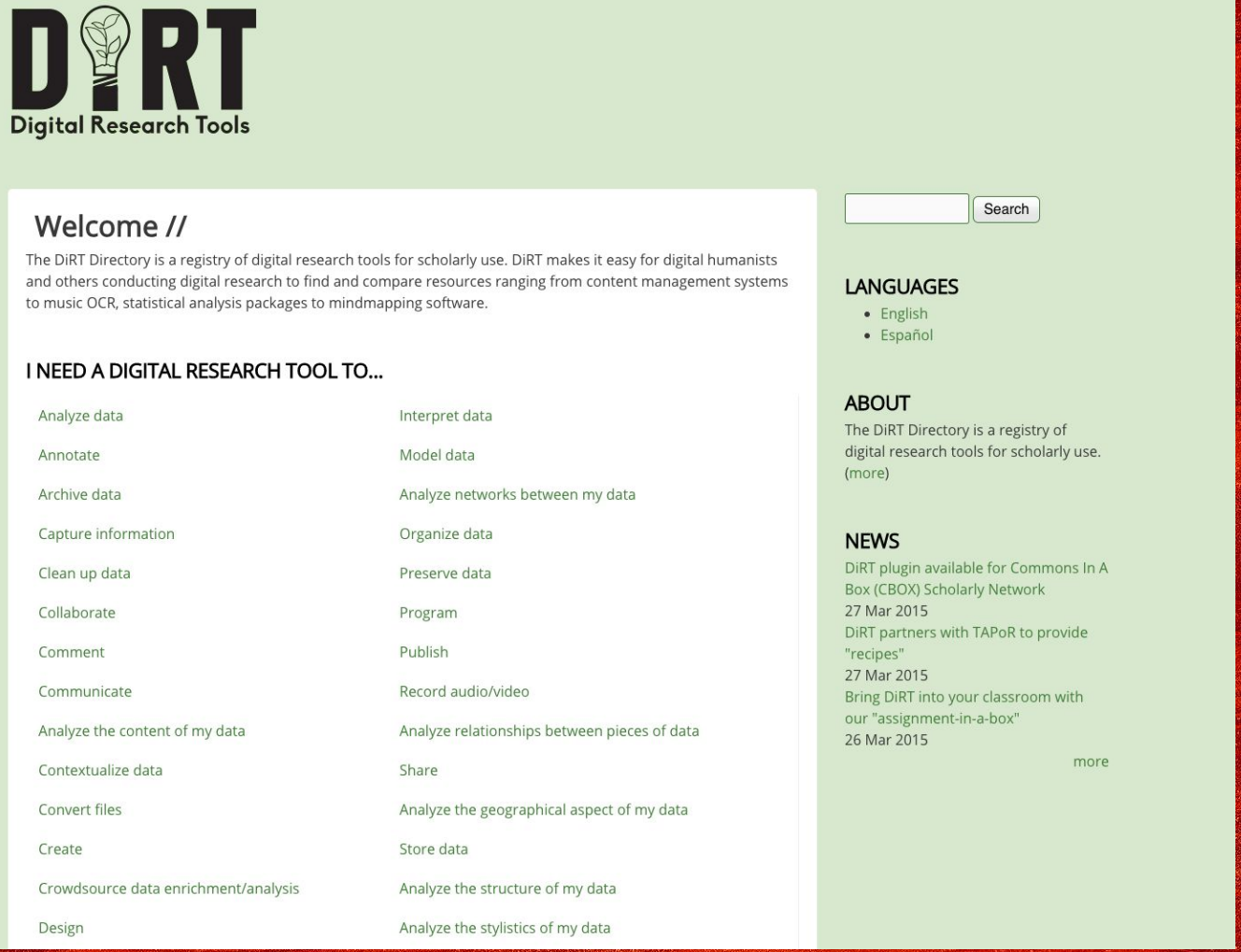


Outputs

Gettin' creative.

dirtdirectory.org

Mellon-funded online directory of computational tools for working with humanities data



Tools:

- AGTK: Annotation Graph Toolkit: "a suite of software components for building tools for annotating linguistic signals, time-series data which documents any kind of linguistic behavior (e.g. audio, video). The internal data structures are based on annotation graphs" (Open source, Java-based/Windows)
- BaseLing: "an online database for linguistic data...a repository of pedagogical exercises for use by students and instructors" (Free, web-based)
- CiteLing: a "master bibliography (*.bib) for all your linguistic needs" (Free, web-based)
- CLaRK: "XML-based system for corpora development"; used for corpora markup, dictionary compilation for human users, and corpora investigation (Free software written in Java)
- CMLaTeX: "a collection of information, links, advice, and bits of code for Linguists who use LaTeX"; also includes downloads for linguistic and LaTeX resources (Free, web-based)
- Computing Optimality with Python: "a tutorial on implementing OT in Python"; includes chunks of Python code and a how-to (Free, web-based)
- ConstraintWiki: "a repository for constraints used in Optimality Theory analyses" (Free, web-based)
- EXMARaLDA: a system for creating, managing and analysing spoken language corpora
- Erculator: a "web-based software that lets you create Optimality Theoretic (OT) tableaux, check their consistency, make inferences about winning candidates and plausible constraint rankings within and across tableaux, explore language typologies, and generate images (png, ps, pdf, Latex) of tableaux for direct inclusion in Word, LaTeX, and other documents" (Free, web-based)
- JGAAP: Java-based,modular, program for textual analysis, text categorization, and authorship attribution
- MonoConc: a "concordance (text searching) program...used in the analysis of English or other texts...also produces wordlists and collocation information" (Commercial, Windows)
- Natural Language Toolkit: "Open source Python modules, linguistic data and documentation for research and development in natural language processing, supporting dozens of NLP tasks, with distributions for Windows, Mac OSX and Linux." (Free, Windows/Mac/Linux)
- Praat: doing phonetics by computer: "a computer program with which you can analyse, synthesize, and manipulate speech, and create high-quality pictures for your articles and thesis" (Open source, cross-platform)
- Saplo: A text analysis API with text recommendations, text filtering, text categorization, automatic tagging, automatic related articles and sentiment analysis. Read more in the text analysis API documentation (Free to try; special offers for researchers and universities are available, web-based).
- Stanford POS Tagger: "a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc." (Open source, cross-platform) [Review]
- TACTWeb: text analysis software that makes TACT TDB databases available on the web to both TACT and non-TACT users (Free, web-based)
- TIGERSearch: a search program that "lets you explore linguistically annotated texts. For example, a lexicographer or terminologist can use TIGERSearch to find out about lexical properties of a word like the collocations the word is used in." (Open source, cross-platform)
- Toolbox (The Field Linguist's Toolbox): "a data management and analysis tool for field linguists. It is especially useful for maintaining lexical data, and for parsing and interlinearizing text, but it can be used to manage virtually any kind of data." (Free, cross-platform)
- Transana: "a computer program that allows researchers to transcribe and analyze large collections of video and audio data" (Commercial; open source, Windows/Mac)
- Transformer: "a software tool for scientists who work with transcribed linguistic data. It addresses conversation analysts, phoneticians, anthropologists, and other social scientists who want to analyze digital audio or video data and language. The Transformer is a program to manage and convert transcribed linguistic and aligned data in a quick, safe, and easy way." (Commercial, Windows)
- Voicewalker: "a transcriber's tool, designed to help you transcribe audio or video recordings. VoiceWalker lets you play back the sound in a controlled way, with the benefit of being able to systematically step (or "walk") through a recording, repeating short segments for a specified number of repetitions, then moving on to the next segment" (Free, Windows)
- WMatrix: "a software tool for corpus analysis and comparison. It provides a web interface to the USAS and CLAWS corpus annotation tools, and standard corpus linguistic methodologies such as frequency lists and concordances. It also extends the keywords method to key grammatical categories and key semantic domains." (Annual subscription, web-based)

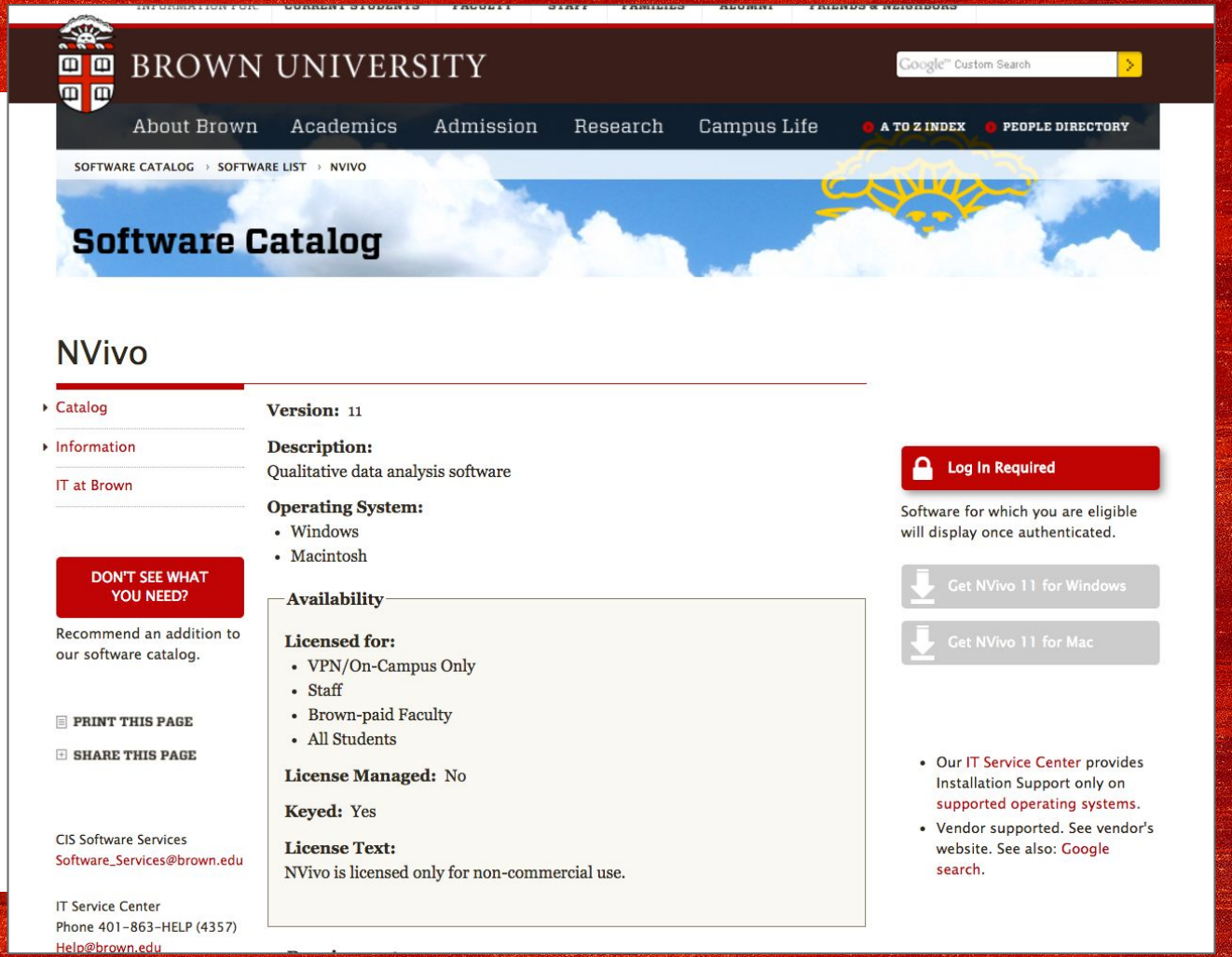
Resources:

- Intute: Linguistics
- Linguist List: Software
- SIL, Linguistics Computing Resources on the Internet
- Linguistic Annotation WiKi

See also:

- Text Analysis Tools
- Transcription Tools

Brown offers tools for analyzing transcripts, audio, and video qualitative data through its nVivo software license. Another popular tool is Atlas.ti.



Ephemera Library

You may want to do this if you have:

- Posters
- Flyers
- Social Media (Tweets, etc.)
- Images
- Correspondance
- Info. graphics
- Business cards

Doster. American Libraries. Jan 2016.

Interested?

Why not Ephemera?

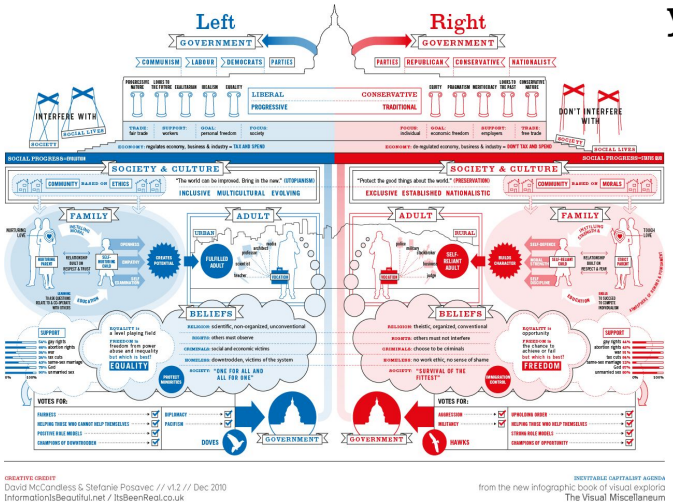
Saving Digital Ephemera

The Ephemera Society of America

Visualizations

Interested?

[Chart.js]
[Datahero]
[Plot.ly]
[Visual.ly]
[Raw]
[Omeka]
[Neatline]
[Exhibit]
[Leaflet]



CREATIVE CREDIT
David McCandless & Stefanie Posavec // v1.2 // Dec 2010
InformationIsBeautiful.net / ItsBeenReal.co.uk

INEVITABLE CAPITALIST AGENDA
from the new infographic book of visual exploria
The Visual Miscellaneum

http://www.informationisbeautiful.net/visualizations/left-vs-right-us/

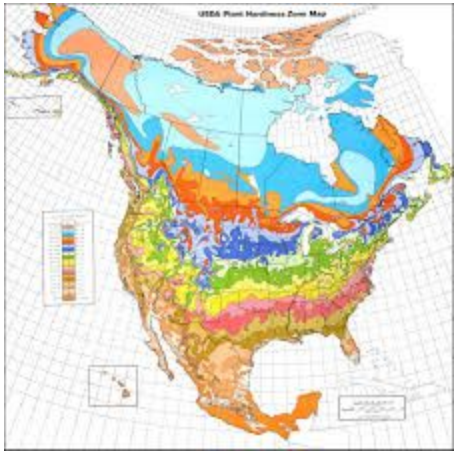
You may want to do this if you have:

- Geo-spacial Data
- GIS
- Locations
- Monuments
- People
- Events
- Textual data
- Statistical data

Mapping

You may want to do this if you have:

- Geo-spacial Data
- GIS
- Locations
- Monuments
- People
- Events



https://goo.gl/v7TQJG

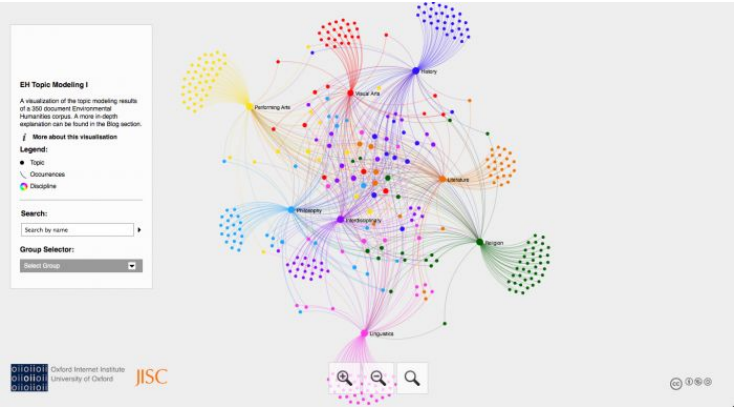
Interested?

Omeka
Neatline
Exhibit
Leaflet

Topic Modeling

You may want to do this if you have:

- Huge quantities of textual data
- A corpus of work
- Translations



Oxford Internet Institute. http://dig-eh.org/dig-eh/TopicModelling/RectangularNodes/. Accessed March 20, 2017.

Interested?

Digital Environmental Humanities

Stanford Topic Modeling Toolbox

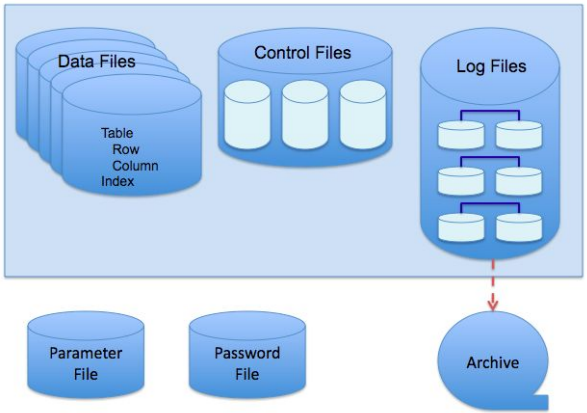
Topic Modeling: A Basic Introduction

Databases

Interested?

Creatly
Navicat
RazorSQL
Vertabelo

Introduction to Database Management Systems



https://goo.gl/DyIg9Y

You may want to do this if you have:

- Ethnographic data
- Textual data
- Narrative data
- Field notes
- Recordings
- Images
- Sound
- Video
- Statistical data

Publishing, Copyright (& left)

Know the types of licenses that restrict your data.

Definition of **copyright:**

- Copyright is a form of protection provided by the laws of the United States (title 17, U.S.Code) to the authors of “original works of authorship,” including literary, dramatic, musical, artistic, and certain other intellectual works. This protection is available to both published and unpublished works. Section 106 of the 1976 Copyright Act generally gives the owner of copyright the exclusive right to do and to authorize others to do the following:

©

From Copyright.gov

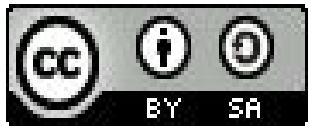
Considerations for copyright

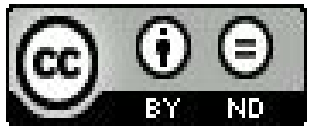
Hannes Grobe
Creative Commons CC-BY-SA-2.5

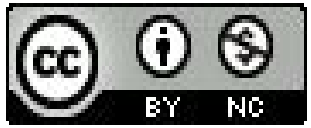
- Who can claim copyright?
- What works are protected?
- How long does the item exist under copyright?
- Can you take pictures of archival items? Can you post/publish them (online)?
- Does your work fall into “fair use”?

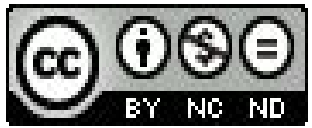
What about **copyleft**?

- derivations
  - licensing passed down
- attribution stacking
- combination
  - prevent the licensed data being combined with data released under a different copyleft licence









Types of **Creative Commons** licensing

Unsure of the license?

Ask the creator.

No response?

Consider "fair use."

(BY) Requires attribution
cite the rights holders as creator

require any derivative open to public use in the same way the rights holders shared with the user

(CC) Creative Commons
use, modify, and redistribute freely on condition that anything derived from it is bound by the same condition

(CC0) Public domain

no restrictions

(NC) No commercial use

remix, tweak, and build upon work non-commercially.

new works must also acknowledge creator and be non-commercial, but they don't have to license derivatives on the same terms.

(ND) No derivatives

may not be remade as derivative; passed along unchanged and in whole

(SA) Share alike

remix, tweak, and build upon work even for commercial purposes

credit creator and license their new creations under the identical terms.

All new works based on yours will carry the same license,

Okay, so what is fair use?

But also consider:

- the purpose of the use
- the amount used
- the nature of the work (creative or factual) and
- the effect your use may have on the market for the copyrighted work.

It may be helpful to think of “fair use” alongside "**transformative use**": *re-purposing of the work rather than simply reproducing it, often in a different context and for a different audience.*

Copyright exceptions & ambiguities

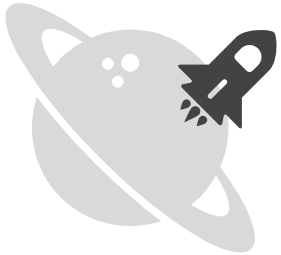
Orphan Works

manifestations of work without the ability to locate the copyright owner after a good-faith qualifying search

Public Domain

a work copyrighted before 1923 is in the Public Domain in the U.S

this means a work can be accessed and used without restrictions as it belongs to the public



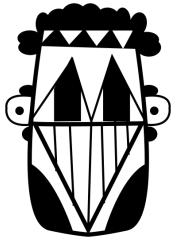
Human Subjects

Considerations for Human Subjects Data

- Government restrictions
- Privacy
- Ethics
- Identifying Information
- Consent

Anonymization

Created by Pete Baker from Noun Project

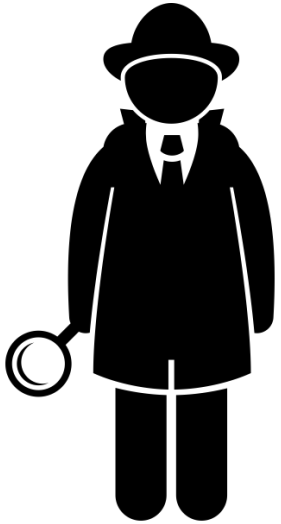


Created by Pete Baker from Noun Project

Created by Pete Baker from Noun Project

De-identification

- Variable Removal
- Top-Coding
- Collapsing or Combining
- Sampling
- Swapping
- Anonymization Scheme



Created by Gan Khoon Lay
from Noun Project

Where can I store Human Subjects Data?

ICPSR

(Virtual) Data Enclave

Brown's Stronghold

REDCap*

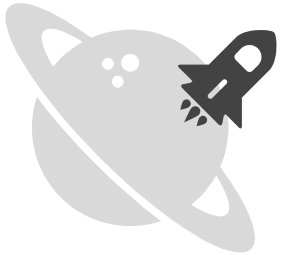
For a directory of repositories, visit: http://www.re3data.org/

Recommended File Formats:

Text: PDF/A, XHTML, HTML, XML, ODF.

Image: TIFF, JPEG2000, GIF.

GIS: SHP, BDF, GEOJSON



Confidentiality Plans

Shh...

What makes something confidential?

Protected Health Information (PHI)

- the individual's past, present or future physical or mental health or condition, the provision of health care to the individual, or
- the past, present, or future payment for the provision of health care to the individual
- the individual's identity or for which there is a reasonable basis to believe it can be used to identify the individual.

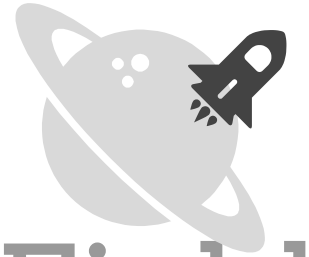
Personally Identifiable Information (PII)

(1) Social Security Number
(2) Driver's license number or California Identification Card number
(3) Account number, credit or debit card number, in combination with any required security code, access code, or password that would permit access to an individual's financial account

Considerations for Confidentiality Plans:

- Addresses how the investigators will work with partners to ensure confidentiality
- Outlines what private information will be collected and for how long it will be stored
- Describes participants' rights in terms of access to their own information (what rights they have to access it)

Please visit Brown's IRB for more information: https://goo.gl/j1sVGX



Post-Fieldwork Considerations

QDR BETA
QUALITATIVE DATA REPOSITORY

Log In Register Contact

Site Search

About Deposit Discover Guidance and Resources Publications MCP

Facilitating use of cutting edge research methods for data generation and analysis

Learn More >>

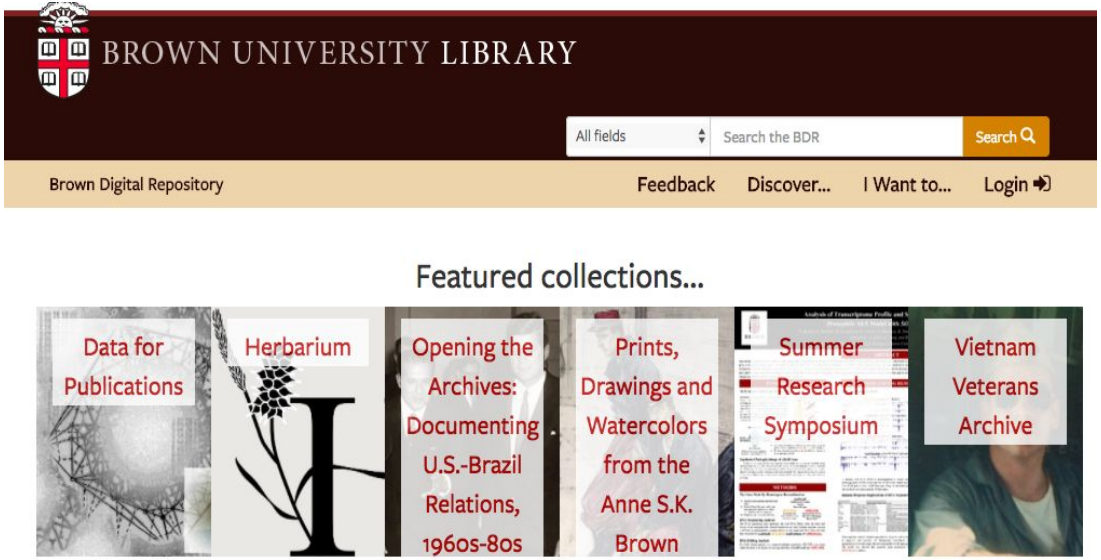
Other publication considerations

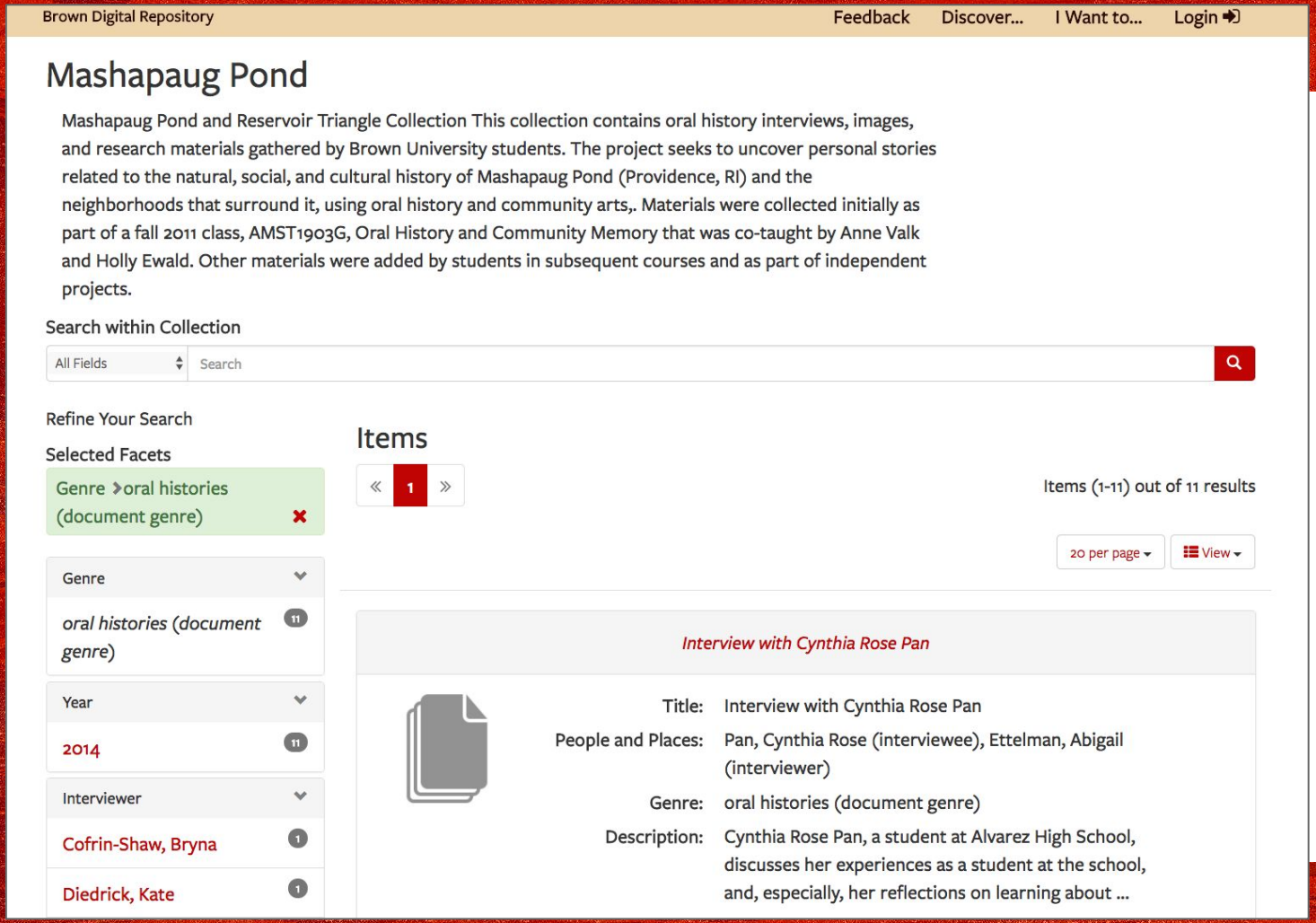
Use the Brown Digital Repository

Digital Object Identifiers (DOIs) for citing digital materials in your dissertation and publications

Embargoes for publication

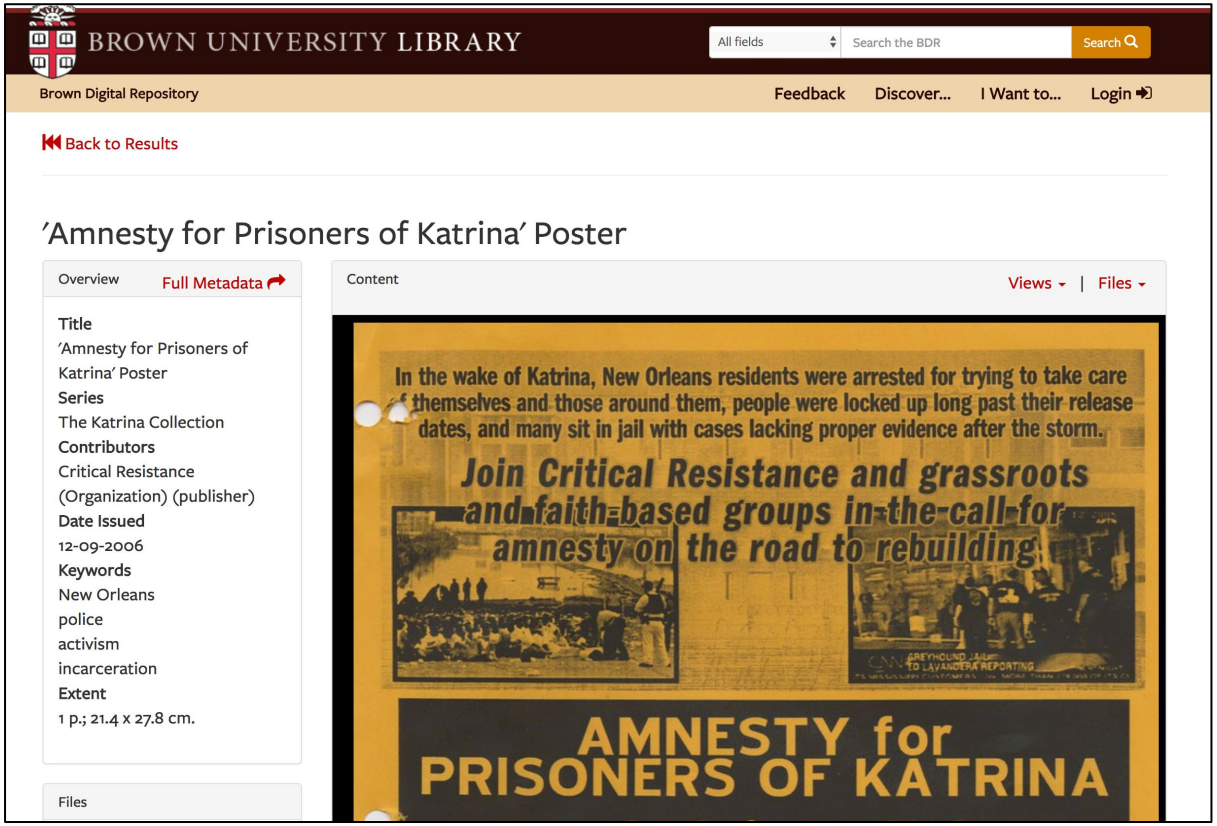
ORCID researcher identification



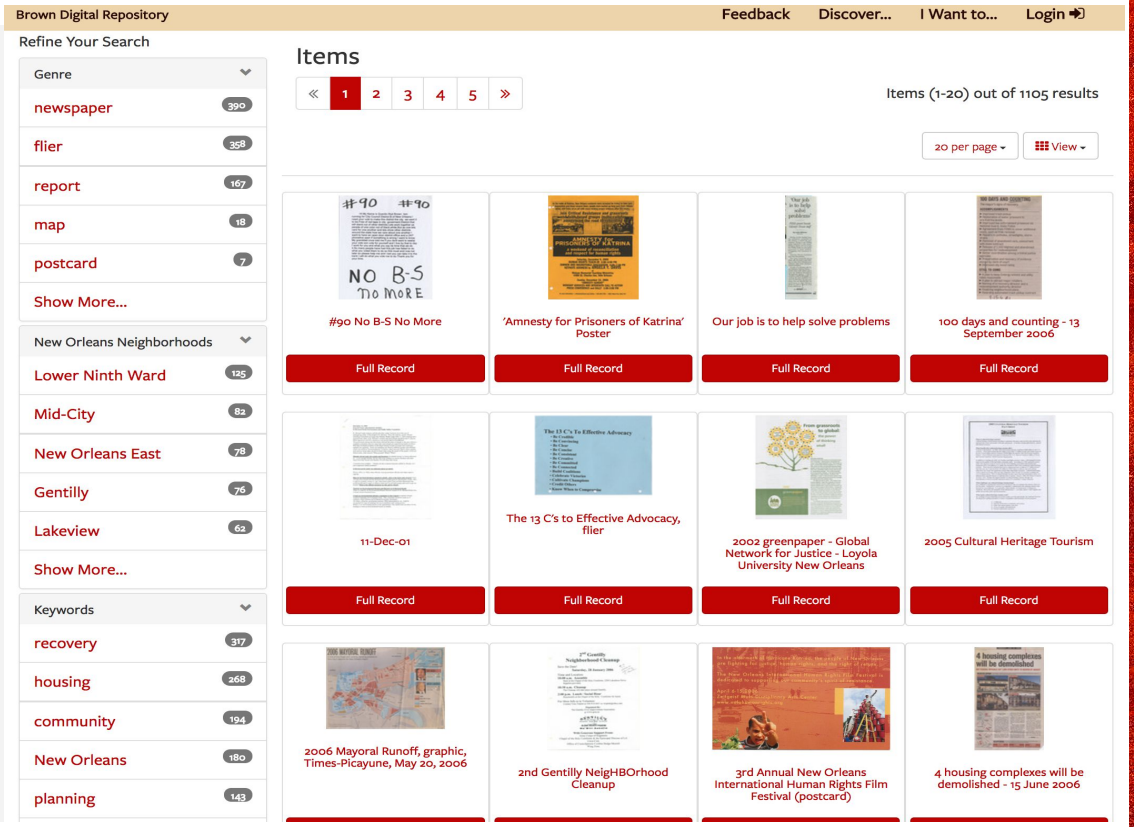


This project contains the audio and transcripts of oral histories within a community in Providence.

In this BDR collection Professor Logan stores the digitized documents that he analyzed studying New Orleans after Hurricane Katrina in 2006.



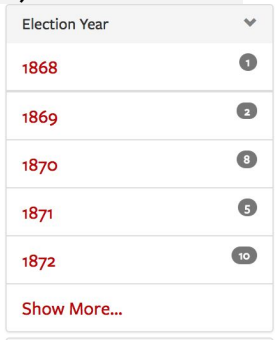
The metadata on the side, called facets, derive from the description that the researcher creates for each of these documents that he analyzed. Important information such as what type of document is it, where did it originate, keywords, etc.

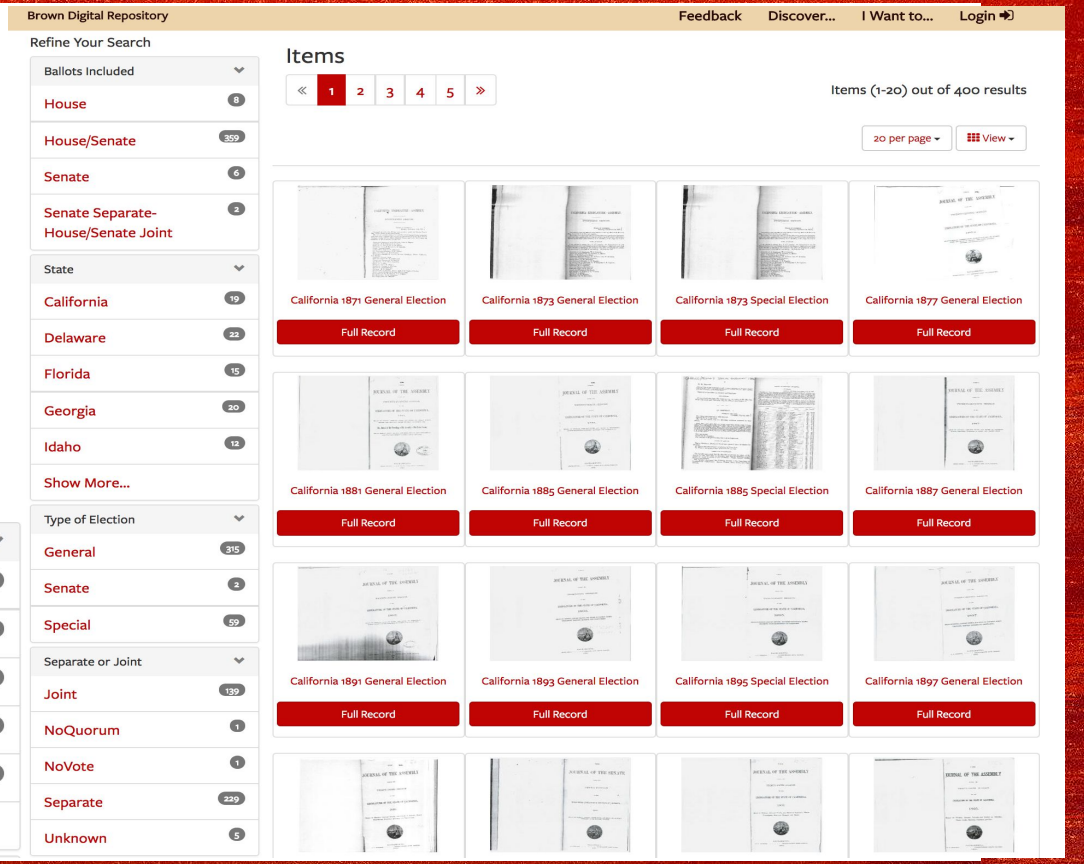


Senate Elections Data Project 1871-1913

In the Senate Elections Data Project 1871-1913, largely funded by a National Science Foundation grant (NSF 0517813; Principal Investigator, Wendy J. Schiller) we have collected the roll call votes from state and house state journals for all identified elections for U.S. senator held in during this time period. We have also collected data on the state legislators who cast these votes, and identified the universe of candidates who sought and won Senate seats during this time period. Where we could, we also collected data on chamber officer votes, and election returns for state legislatures, as well as a large collection of primary source newspaper articles about these elections. There has never been a systematic account of how indirect elections worked across all states, especially with respect to how candidates were nominated and elected, the nature of the conflict over these seats, and the role and strength of party organizations in influencing the outcome. It is our hope that making this data public will encourage others to ask and answer other important questions about electoral systems, legislative behavior, American history, and the efficacy of changing the Constitution to produce more responsive government.

In this BDR collection, Professor Schiller has created a digital library of her sources underlying her last book on the history of U.S. Senate elections in the late nineteenth and early twentieth century. Note again that that the quality of her description (metadata) allows her to navigate through these sources, such as which state, type of election, year of election, etc.





Professor Houston created a digital library of photographs from his excavation in Piedras Negras. This is a great way to preserve these files and cite an image in a publication by including the link.

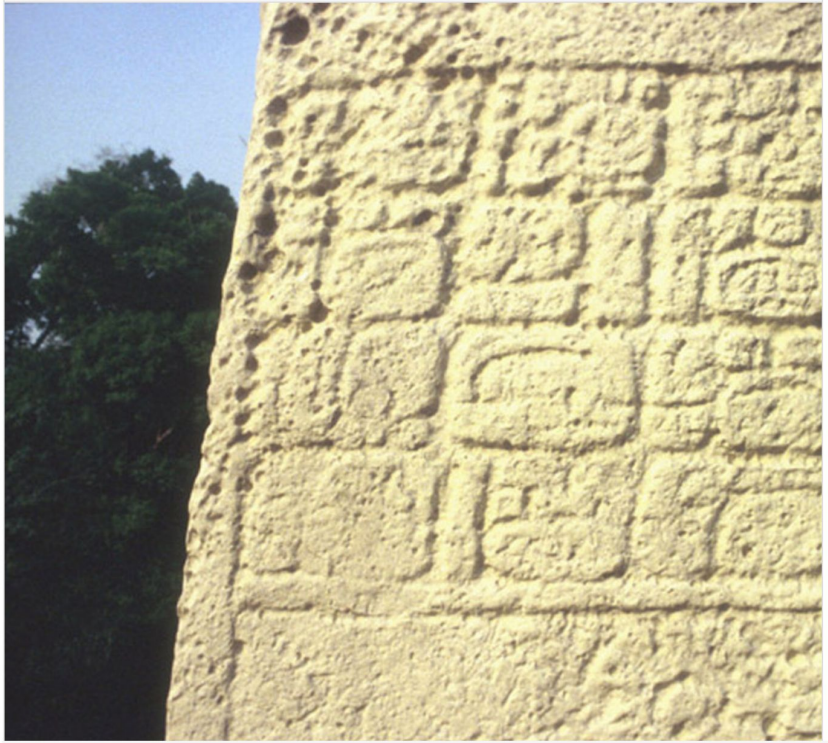
close look at glyphs on a stela

Overview Full Metadata

Title
close look at glyphs on a stela
Contributors
Houston, Stephen D.
Date Created
06-19-1905
Keywords
Archaeology
Mayas
Notes
Original Medium: 35mm color
Image Manipulations from Original: image sizing, brightness/contrast

Files
METS
MODS
jpg

Content Views | Files



Citation

"close look at glyphs on a stela" (1905). *Piedras Negras Images.* Brown Digital Repository. Brown University Library. https://repository.library.brown.edu/studio/item/bdr:222445/

Thanks!

Any questions?

You can find us at

- @AndyDrewCreamer or @simas_sam
- andrew_creamer@brown.edu or samuel_simas@uri.edu