Storing, Backing Up and Archiving Data

Jean Aroom, Clinton Heider and Lisa Spiro

Objectives for This Session

- Know options for storing, backing up, sharing and archiving your data.
- Understand best practices for protecting your data.

Data Storage Definition

- The media (optical or magnetic) to which you save your data files and software.
- All storage media are vulnerable to risk and obsolescence.
- Storage media should be evaluated and updated every 2-5 years.

New England Collaborative Data Management Curriculum

Data Storage Considerations

- Location (Internal/External HD, Network, Remote)
- Disk size or storage quota
- Computing performance
- Accessibility

Data Backup Definition

- Allows you to *restore* your data if original data is lost or damaged due to:
  - Hardware or software malfunction
  - Environmental disaster (fire, flood)
  - Theft
  - Unauthorized access

New England Collaborative Data Management Curriculum

3-2-1 Backup Rule

3 Save 3 copies of your data.

2 Use 2 types of storage.

1 Keep 1 remote copy.

Data Backup Considerations

- Location (On-site, off-site)
- Procedure (Full, differential, incremental, mirror)
- Frequency (Hourly, daily, weekly, monthly)
- Retention (Months, years)
- Performance

TEST YOUR BACKUP PLAN!

Data Backup Summary

Backup type	Backed up	Backup time	Restore time	Storage space
Full/snapshot	All data	Slowest	Fast	High
Differential	All data since last full	Moderate	Moderate	Moderate
Incremental	Only new/ modified files	Fast	Slowest	Lowest
Mirror	Only new/ modified files	Fastest	Fastest	Highest

Overview of Data Storage, Backup and Sharing Options at Rice

Network Storage

- **storage.rice.edu** - U: drive, departmental shares
- **Research Data Facility (RDF)** - larger scale storage for research projects

Backup Options

- **storage.rice.edu** backups/snapshots
- **Crash Plan** for Rice workstations

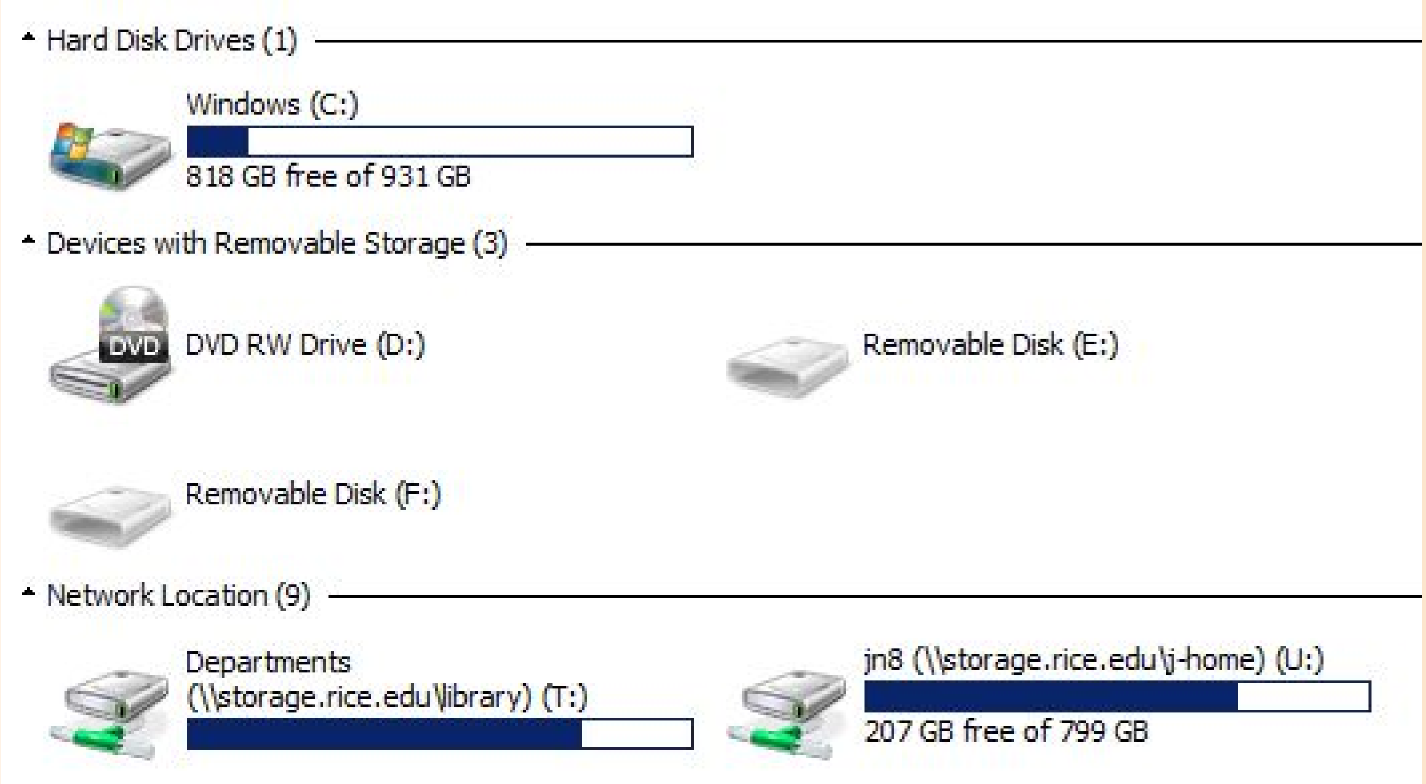
Data Sharing/Collaboration Tools - Google Drive, Rice Box, Globus Connect

Options for faculty/ staff: https://kb.rice.edu/page.php?id=70762

Options for students: https://kb.rice.edu/page.php?id=65636

Storage: storage.rice.edu

- Location: Networked
- Storage quotas
    - Undergraduates: 2 GB
    - Graduates, Staff, Faculty: 5 GB
    - Colleges, Depts, Centers, Institutes: 40 GB
- Performance - Subject to network
- Accessibility
    - NetID folder: Private, not shared
    - Groups: Any Rice NetID holder by request



Hard Disk Drives (1)

Windows (C:)
818 GB free of 931 GB

Devices with Removable Storage (3)

DVD RW Drive (D:)

Removable Disk (E:)

Removable Disk (F:)

Network Location (9)

Departments (\\storage.rice.edu\library) (T:)

jn8 (\\storage.rice.edu\j-home) (U:)
207 GB free of 799 GB

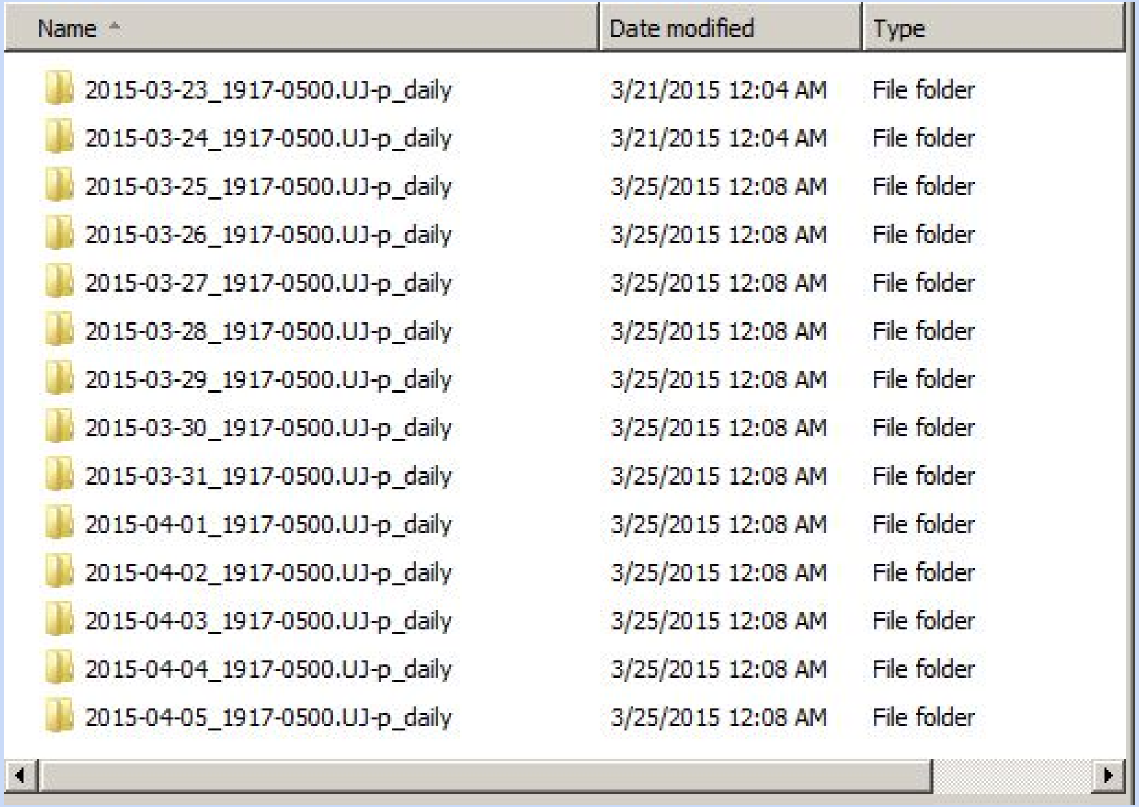
\\storage.rice.edu

Storage: Research Data Facility

- Location: On Site (Rice PDC) network data shares
- Storage quotas
  - 500GB per researcher
  - Additional storage available with cost recovery
  - Cost below $100/TB/year, prorated monthly by use
- Performance - Subject to network
- Accessibility
  - Based on NetID and ADRICE security groups
  - Can be shared to multiple users in a research group

Backup: storage.rice.edu

- Location: On-site
- Procedure: Full replication
- Frequency: Daily
- Retention
  - Personal access: 2 weeks
  - Request IT restoration: 6 months

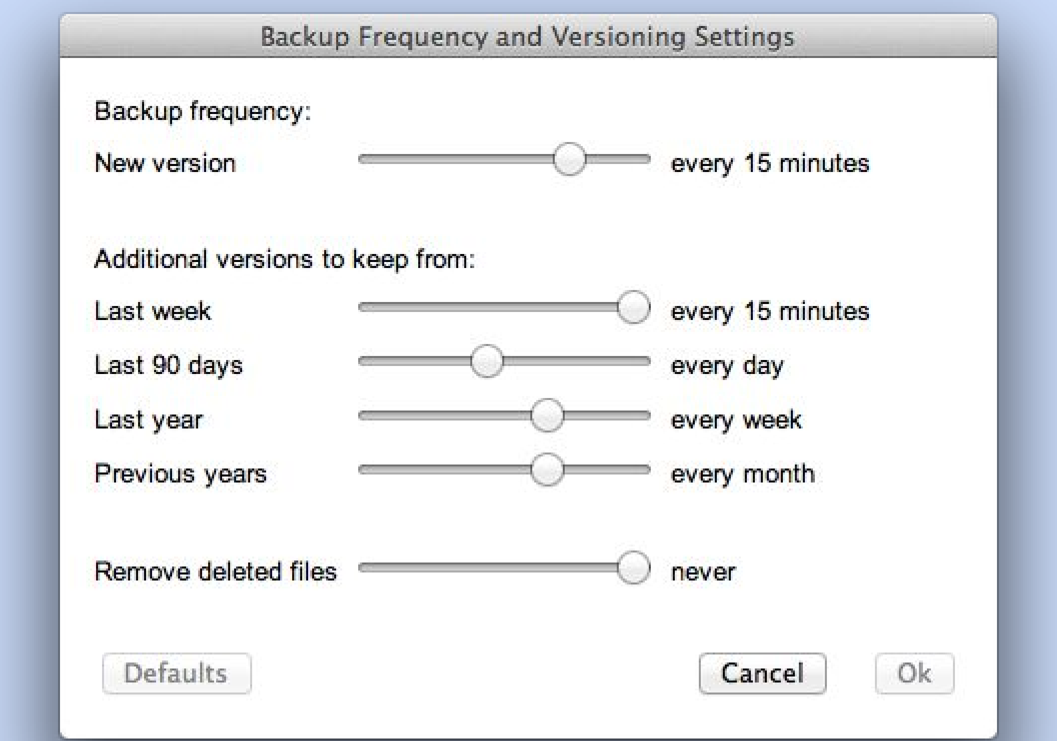


Name	Date modified	Type
2015-03-23_1917-0500.UJ-p_daily	3/21/2015 12:04 AM	File folder
2015-03-24_1917-0500.UJ-p_daily	3/21/2015 12:04 AM	File folder
2015-03-25_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-26_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-27_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-28_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-29_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-30_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-03-31_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-01_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-02_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-03_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-04_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder
2015-04-05_1917-0500.UJ-p_daily	3/25/2015 12:08 AM	File folder

\\storage.rice.edu\?-home\~snapshot

Backup: CrashPlan

- Availability: Rice-owned computers
- Cost: $82.56/year/person (up to 4 devices)
- Location: Off-site cloud storage
- Procedure: Incremental
- Frequency: Adjustable up to every minute
- Retention: Adjustable up to forever



Backup Frequency and Versioning Settings

Backup frequency:

New version: every 15 minutes

Additional versions to keep from:

Last week: every 15 minutes

Last 90 days: every day

Last year: every week

Previous years: every month

Remove deleted files: never

CrashPlan PROe or crashplan.rice.edu

Sharing: Google Drive

- Unlimited storage for low risk data
- Can be used for collaboration within Rice
- Integrates nicely with G-suite productivity apps
- Files aren’t local and performance is limited
- No provisions for retention of orphaned data
- Accessibility
    - Login to G-Suite apps with your Rice NetID

Sharing: Rice Box

- Web based file sharing tool similar to Dropbox
- Approved by Rice for sharing secure data
- Accessibility
    - Rice NetID
    - Share folders with Rice colleagues or external collaborators
    - Add emails of external collaborators to a folder and send invitations

Sharing: Globus Connect

- Widely used service for large data exchange between participating institutions
- Can be used in our HPC environment or from your desktop with Globus Connect Personal
- Accessibility
    - Contact CRC to be added to license
    - Arrange for access to peer institution end points

Product	Use	Location	Quota	Performance	Accessibility
Storage	S/B	Rice Data Center	2-5-40 GB	Network	NetID
Google Drive	S/C	Global Cloud	Unlimited	Internet	NetID & External
RDF	S/B	Rice Data Center	500GB free	Network	NetID
Rice Box	S/C	US Cloud	Unlimited	Internet	NetID & External
CrashPlan	B	Off-site cloud	Unlimited	Internet	Your NetID

Data Security

- Confidential (SSN, CC#, DL#)
  - Financial records
  - Health records
  - Education records

- Sensitive (Birth date, address, emergency contact, EID/SID)

Security Classification	Rice On-Site Most Secure	Rice Cloud Contract Semi-Secure
Low Risk (Public Data)	CampusPress, RDF	Google Drive
Moderate Risk (Sensitive Data)	RDF	Rice Box
High Risk (Confidential Data)	Storage Confluence	Rice Box CrashPlan
High Risk (Regulated Data)	Storage	CrashPlan

Data Archiving Definition

- Provides a final version of your data
- Stored for the long-term

Data Archiving Considerations

- Location
- File formats
- Responsibility
- Accessibility

Why Archive Your Data with a Data Repository?

- Conform to publisher or funder requirements
- Get cited
    - “studies that made data available in a public repository received 9% … more citations than similar studies for which the data was not made available.” (Piowowar & Vision, 2013)
- Promote future research

Data Archiving Options

Public Repositories:

- Discipline based repository
- General data repository (e.g. FigShare)
- Rice Digital Scholarship Archive

Private Approaches:

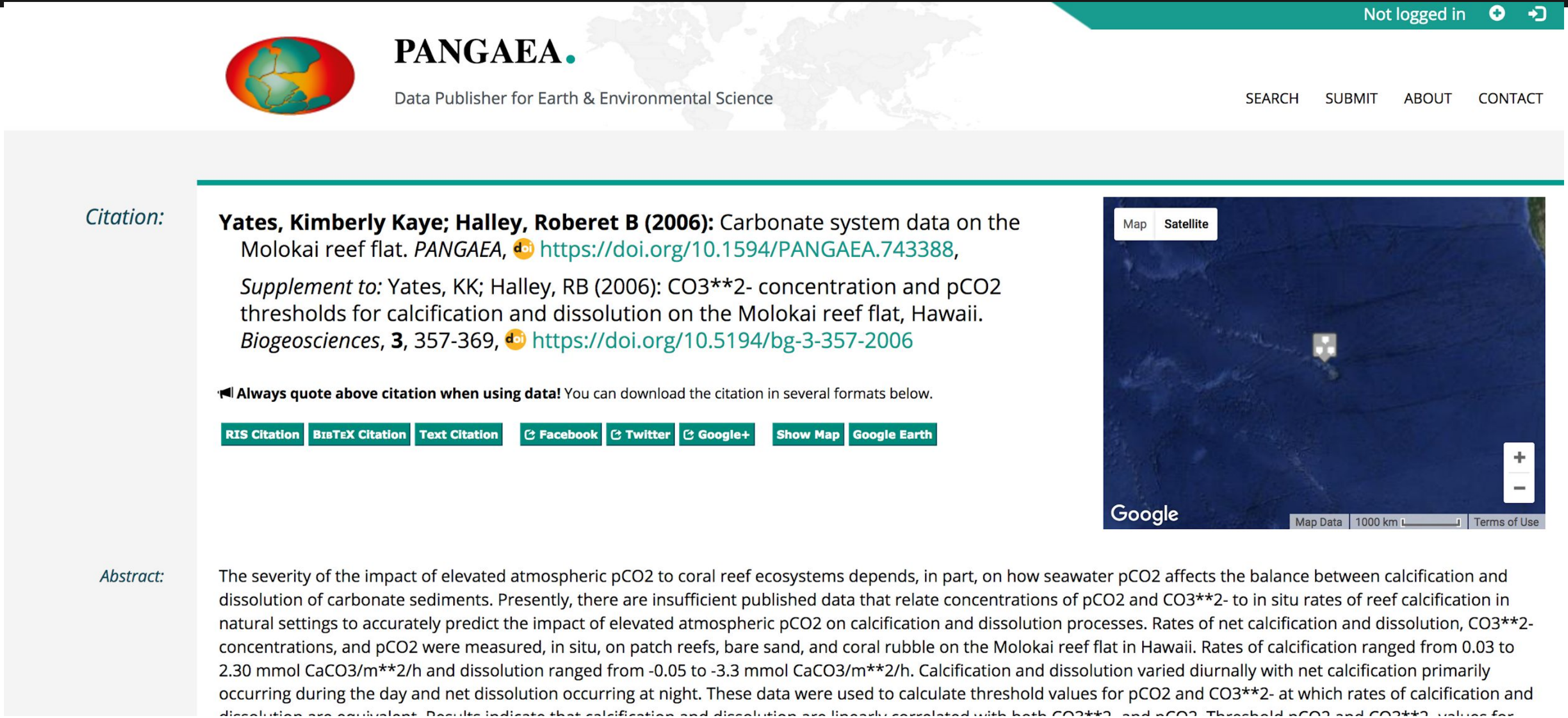
- Long-term storage (redundant)

Finding a repository

Consult lists and directories of data repositories:

- Nature, “Recommended Data Repositories”: https://www.nature.com/sdata/policies/repositories
- PLOS Guide: http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories
- Re3data: http://www.re3data.org/

Share Your Data through A Disciplinary Repository: Pangea



Citation:

Yates, Kimberly Kaye; Halley, Roberet B (2006): Carbonate system data on the Molokai reef flat. *PANGAEA*, https://doi.org/10.1594/PANGAEA.743388,

Supplement to: Yates, KK; Halley, RB (2006): CO3**2- concentration and pCO2 thresholds for calcification and dissolution on the Molokai reef flat, Hawaii. *Biogeosciences*, **3**, 357-369, https://doi.org/10.5194/bg-3-357-2006

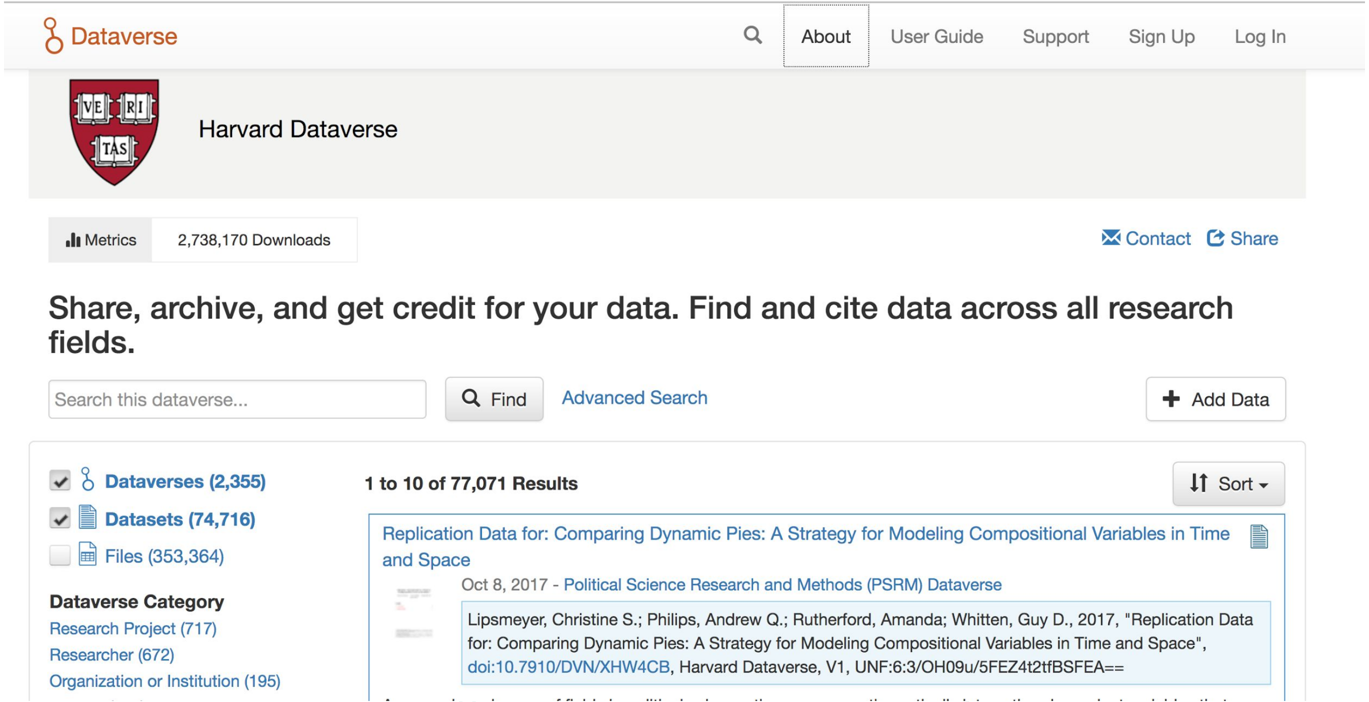
Always quote above citation when using data! You can download the citation in several formats below.

Abstract:

The severity of the impact of elevated atmospheric pCO2 to coral reef ecosystems depends, in part, on how seawater pCO2 affects the balance between calcification and dissolution of carbonate sediments. Presently, there are insufficient published data that relate concentrations of pCO2 and CO3**2- to in situ rates of reef calcification in natural settings to accurately predict the impact of elevated atmospheric pCO2 on calcification and dissolution processes. Rates of net calcification and dissolution, CO3**2- concentrations, and pCO2 were measured, in situ, on patch reefs, bare sand, and coral rubble on the Molokai reef flat in Hawaii. Rates of calcification ranged from 0.03 to 2.30 mmol CaCO3/m**2/h and dissolution ranged from -0.05 to -3.3 mmol CaCO3/m**2/h. Calcification and dissolution varied diurnally with net calcification primarily occurring during the day and net dissolution occurring at night. These data were used to calculate threshold values for pCO2 and CO3**2- at which rates of calcification and dissolution are equivalent. Results indicate that calcification and dissolution are linearly correlated with both CO3**2- and pCO2. Threshold pCO2 and CO3**2- values for

https://doi.pangaea.de/10.1594/PANGAEA.743388

Harvard Dataverse



Harvard Dataverse

2,738,170 Downloads

Share, archive, and get credit for your data. Find and cite data across all research fields.

Search this dataverse...

☑ Dataverses (2,355)
☑ Datasets (74,716)
□ Files (353,364)

Dataverse Category

Research Project (717)
Researcher (672)
Organization or Institution (195)

1 to 10 of 77,071 Results

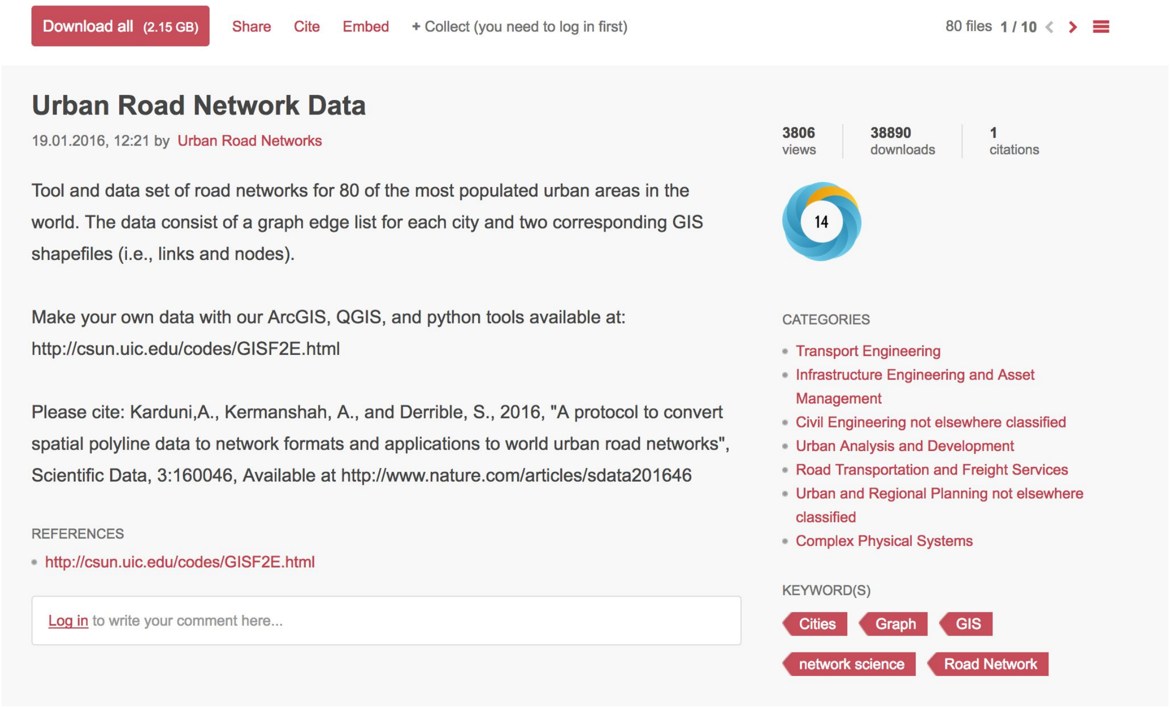
Replication Data for: Comparing Dynamic Pies: A Strategy for Modeling Compositional Variables in Time and Space

Oct 8, 2017 - Political Science Research and Methods (PSRM) Dataverse

Lipsmeyer, Christine S.; Philips, Andrew Q.; Rutherford, Amanda; Whitten, Guy D., 2017, "Replication Data for: Comparing Dynamic Pies: A Strategy for Modeling Compositional Variables in Time and Space", doi:10.7910/DVN/XHW4CB, Harvard Dataverse, V1, UNF:6:3/OH09u/5FEZ4t2tfBSFEA==

https://dataverse.harvard.edu/

Figshare



Urban Road Network Data

19.01.2016, 12:21 by Urban Road Networks

3806	38890	1
views	downloads	citations

Tool and data set of road networks for 80 of the most populated urban areas in the world. The data consist of a graph edge list for each city and two corresponding GIS shapefiles (i.e., links and nodes).

Make your own data with our ArcGIS, QGIS, and python tools available at: http://csun.uic.edu/codes/GISF2E.html

Please cite: Karduni,A., Kermanshah, A., and Derrible, S., 2016, "A protocol to convert spatial polyline data to network formats and applications to world urban road networks", Scientific Data, 3:160046, Available at http://www.nature.com/articles/sdata201646

REFERENCES

- http://csun.uic.edu/codes/GISF2E.html

Log in to write your comment here...

CATEGORIES

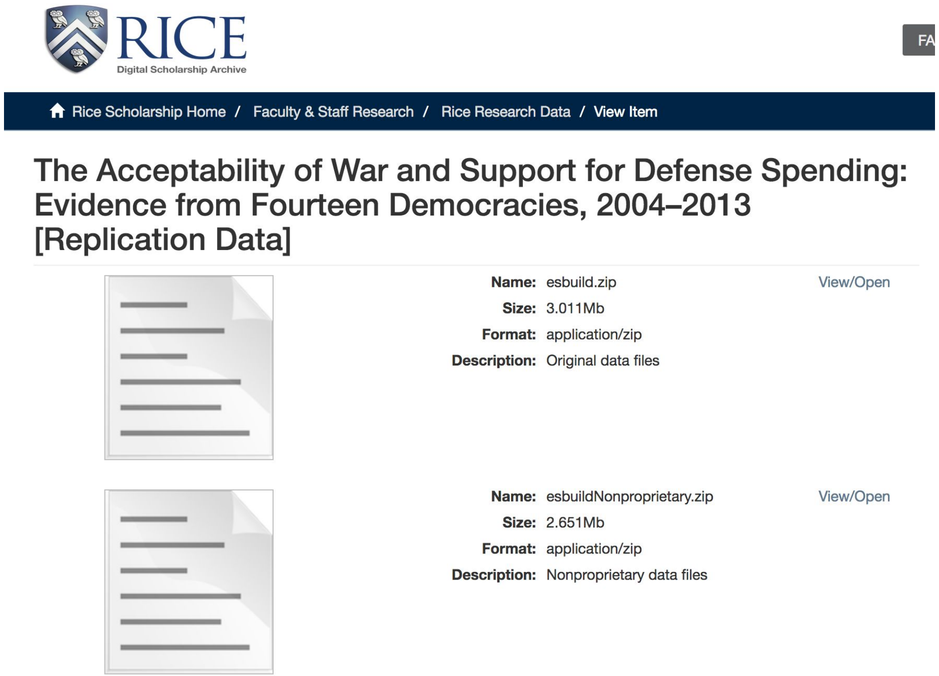
- Transport Engineering
- Infrastructure Engineering and Asset Management
- Civil Engineering not elsewhere classified
- Urban Analysis and Development
- Road Transportation and Freight Services
- Urban and Regional Planning not elsewhere classified
- Complex Physical Systems

KEYWORD(S)

Cities Graph GIS network science Road Network

https://figshare.com/

Rice Data Sharing Option: Rice Digital Scholarship Archive


RICE
Digital Scholarship Archive
FA
Rice Scholarship Home / Faculty & Staff Research / Rice Research Data / View Item

The Acceptability of War and Support for Defense Spending: Evidence from Fourteen Democracies, 2004–2013 [Replication Data]

Name: esbuild.zip View/Open
Size: 3.011Mb
Format: application/zip
Description: Original data files

Name: esbuildNonproprietary.zip View/Open
Size: 2.651Mb
Format: application/zip
Description: Nonproprietary data files

https://scholarship.rice.edu/

How to Set Yourself Up to Archive Your Data

- Before sharing, ensure that confidentiality is protected and that there are no copyright concerns.
- Document your data so that others understand it.
- Organize your data
- Provide the metadata required by the repository
- Get your data into the appropriate format (ideally a non-proprietary format like csv or txt)
- Provide metadata
- Aim for networked storage rather than device-dependent

Example of submission requirements: Pangea

Documentation

--explain abbreviations

--provide units for parameters

Metadata:

-- position (geographic)

--citation of journal article

Format:

--excel or tab-delimited text files for tables

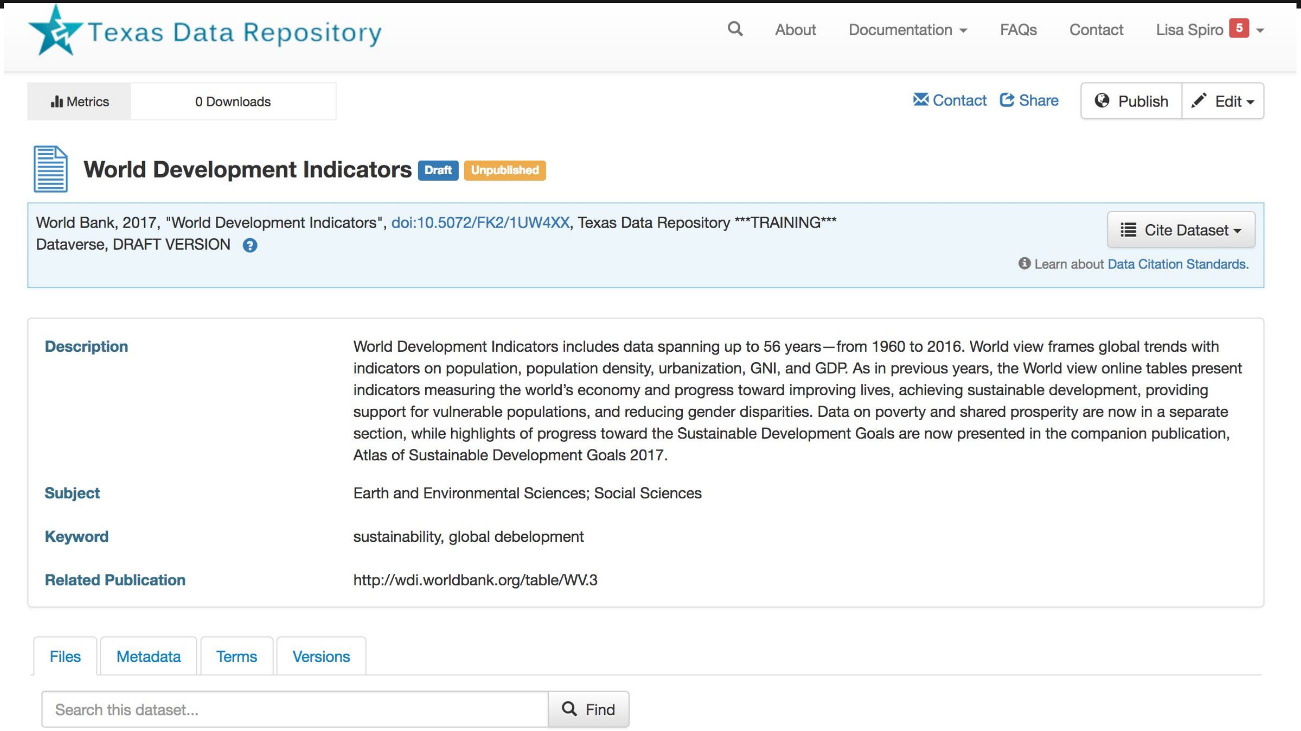
Questions to Ask in Evaluating a Data Repository

1. How well will the data be preserved? How stable is the repository?
2. What kind of reputation does the archive have in the community?
3. Does the repository facilitate citation of the data?
4. Does the repository allow you to describe the data fully and make it discoverable?
5. Are there curators who can help to deposit the data?
6. What are the costs of deposit, if any?

Data Archiving Caveats

- Do not share confidential data (unless it has been de-identified and approved through IRB).
- Consult with your collaborators before publishing data.
- It may be possible to embargo data so that it is not available until the related publication is released.

Offer Your Input: Texas Data Repository



World Development Indicators Draft Unpublished

World Bank, 2017, "World Development Indicators", doi:10.5072/FK2/1UW4XX, Texas Data Repository ***TRAINING*** Dataverse, DRAFT VERSION

Description	World Development Indicators includes data spanning up to 56 years—from 1960 to 2016. World view frames global trends with indicators on population, population density, urbanization, GNI, and GDP. As in previous years, the World view online tables present indicators measuring the world's economy and progress toward improving lives, achieving sustainable development, providing support for vulnerable populations, and reducing gender disparities. Data on poverty and shared prosperity are now in a separate section, while highlights of progress toward the Sustainable Development Goals are now presented in the companion publication, Atlas of Sustainable Development Goals 2017.
Subject	Earth and Environmental Sciences; Social Sciences
Keyword	sustainability, global debelopment
Related Publication	http://wdi.worldbank.org/table/WV.3

https://data.tdl.org/

Resources

- DataOne Primer on Data Management, https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf
- Dataverse, *Data Management Plans*, http://best-practices.dataverse.org/data-management/
- ICPSR *Guide to Social Science Data Preparation and Archiving,* http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/
- Svend Juul et al, “Take good care of your data,” http://www.epidata.dk/downloads/takecare.pdf
- UK Data Archive, *Managing and Sharing Data: Best Practices for Researchers*, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Thanks!

Please contact researchdata@rice.edu or with any questions.

Visit us online at http://researchdata.rice.edu/.

Help us shape future workshops! Please complete this evaluation form: https://goo.gl/forms/4kOO9G7Hqrdi79hU2