Tools and Resources for Data Curation

Stephen Abrams Perry Willett UC Curation Center / California Digital Library

> Summer Institute June 2014



Agenda

- Who we are
- Data curation, publication, and sharing
- Tools to help you
 - DMPTool
 - DataUp
 - Dash
 - WAS
- Summary
- Discussion









Who we are









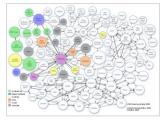
UNIVERSITY OF CALIFORNIA

"To support the University of California community's pursuit of scholarship and ... public service mission"



Data curation, publication, and sharing

- Increasingly, a requirement for funding and publication
- Transparency \leftrightarrow trust
- Reduce needless duplication of effort
- Leverage prior investments
- Expand the reach of your research, and get credit for it
- Good for science, good for scientists



www.flickr.com/photos/_after8_/4052028795



berkeley.edu/teach



www.flickr.com/photos/infocux/8450190120

Data curation, publication, and sharing

- Create/acquire a dataset in a form that is inherently preservable and (re)usable
- Describe the dataset in scientifically-meaningful ways
- Give the dataset a unique identifier for persistent citation
- License the dataset under CC0 OFFICE or CC-BY O
- Deposit the dataset in a (non-commercial) repository where it will receive pro-active curation management
- Expose the dataset for harvesting by abstracting/ indexing services and search engines

DMPTool

• "Fulfill institutional and funder mandates"

	OMPTool		da.			
Home	DMP Requirements	Public DMPs	News Help	Contact Us	About 🔻	🖵 Log In
			1 de la			
Da	ata Mana	ageme	ent			
Pla	anning T	ool		-1-1-		
Creat	te, review, and	share				
	management p utional and fund			Get St	arted	



Build your Data Management Plan

dmptool.org blog.dmptool.org github.com/CDLUC3/dmptool/wiki

DMPTool

- Free and open for all
- Hosted by CDL, with code released as open source
- Supports data management requirements for NSF, NIH, NEH, NOAA, IMLS, and other federal agencies and private funders
- New version released on May 29
- Developed by a partnership of universities, museums, and researchers, with support from Sloan Foundation and IMLS



Build your Data Management Plan

dmptool.org blog.dmptool.org github.com/CDLUC3/dmptool/wiki

DMPTool

- In addition to fulfilling external requirements, the DMPTool provides:
 - Framework to plan for management of research data
 - Comprehensive list of issues involved with data management best practices
 - Information about local resources and services: repositories, workshops, consultation services, etc.
 - Community of stakeholders: researchers, lab managers, IT specialists, archivists, grant administrators, funding agencies



Build your Data Management Plan

dmptool.org blog.dmptool.org github.com/CDLUC3/dmptool/wiki

DataUp

• "Curation for tabular datasets"





dataup.org dataup.cdlib.org

DataUp

• Excel is often the database of choice for research

*DATAUP			Welcome, Stephen Abrams Sign O
My Files view All 🔹	File Name 🔹	🖲 Post 😢 Download 🛞 Dele	te
	TLD_IDs,csv-xlsx Repository	Temp_samples.xlsx Test Repository	
File Name AllFTLD_IDs.csv	Repository N ※ Test F	lame Repository	
File Extension	Identifier		
xlsx	ark:/9999	9/fk4xs86qw	
File Size	Citation		
14.72 КВ		Stephen(2014): AllFTLD_IDs. ams, Stephen. ark:/999999/fk	
Published Date	Status		
3/5/2014 22:20:2	Posted		



dataup.org dataup.cdlib.org

DataUp

- Drag-and-drop data upload
- Opportunity to add descriptive metadata
- Assignment of persistent identifier / generation of persistent citation
- Best practices check
- Packaging and submission to ONE*Share* repository





dataup.org dataup.cdlib.org

BITSS Summer Institute



DataCite









• "Long-term identifiers made easy"

Kong-term id	EZI[entifiers made	easy					Login Contac	: Us
Home	Manage	IDs Create IDs	Lookup ID	Demo				
Why EZID?	Understand	ding Identifiers Pricing I	Documentation C	utreach Who	's using EZID?			
 create store updat use E choos 	identifiers for a citation metada e current URL I ZID's programn e from a variety	nakes it easy to create anything: texts, data, bones, t ata for identifiers in a variety of ocations so citation links are ning interface for automated o y of persistent identifiers, inclu and contact us for more infor	terms, etc. normats never broken peration at scale uding ARKs and Data					
Latest nev	vs from EZID	EZID Service Update: April	2014				<mark>តា</mark> RS	s
		EZID is a service of t	he University of California	Curation Center of	the California Digital Libra	ry		
		Copyright © 2009-2014 The	Regents of the University	of California	About Us Help Priv	acy Policy		



ezid.cdlib.org

• "Long-term identifiers made easy"

Long-term identifiers made easy	Login Contact Us					
Home Manage IDs Create IDs Lookup ID Demo						
Why EZID? Understanding Identifiers Pricing Documentation Outreach Who's using EZID?						
 EZID (easy-eye-dee) makes it easy to create & manage unique, long-term identifiers create identifiers for anything: texts, data, bones, terms, etc. store citation metadata for identifiers in a variety of formats update current URL locations so citation links are never broken use EZID's programming interface for automated operation at scale choose from a variety of persistent identifiers, including ARKs and DataCite DOIs try out EZID anytime and contact us for more information 						
Latest news from EZID Service Update: April 2014	🔊 RSS					
EZID is a service of the University of California Curation Center of the California Digital Library Copyright © 2009-2014 The Regents of the University of California About Us Help Privacy Policy						





• "Long-term identifiers made easy"

Kong-term id	EZID Jentifiers made easy	Login Contact Us
Home	Manage IDs Create IDs Lookup ID Demo	
EZID (eas • create • store • update • use E • choos	by Understanding Identifiers Pricing Docume Agreye-dee) makes it easy to create & man a identifiers for anything: texts, data, bones, terms, etc. citation metadata for identifiers in a variety of formats e current URL locations so citation links are never broken IZID's programming Interface for automated operation at scale se from a variety of persistent identifiers, including ARKs and DataCite DOIs t EZID anytime and contact us for more information)
Latest nev	ws from EZID Service Update: April 2014	S RSS
	EZID is a service of the University of California Curation Center of the California Digital Library Copyright © 2009-2014 The Regents of the University of California About Us Help Privacy Policy	



ezid.cdlib.org

• "Long-term identifiers made easy"

KEZIC) sy		Login Contact Us
Home Manage ID:	s Create IDs Looki	up ID Demo	
Why EZID? Understandir	OLAC R oai:cla.berk	ecord eley.edu:Bloomfield.003	DOI for persistent citation and bi-directional linking
• create identifiers for an	Metadata		between publications and
 store citation metadata update current URL loc use EZID's programmir 	Title: Contributor (researcher):	[Eastern Ojibwa notebook #14] Bloomfield, Leonard [undated]	underlying data
 use E210's programmin choose from a variety c 	Description:	Digital images of index to McGregor field Original dated 1940 or 1941 according t	d no
 try out EZID anytime ar 		56 digital images Bloomfield.003.001	ger mager and p
Latest news from EZID	Identifier (URI); Is Format Of (URI).	http://cla.horkoloy.odu/itom/147 http://dx.doi.org/doi:10.7297/X2BV7DJ2	
	Rights:		age Archive are the physical property of their respective physical copyright, belong to item creators or their legal heirs and assigns.
		Survey of California and Other Indian La	anguages
	Subject: Subject (ISO639):	Eastern Ojibwa language	
	Type (DCMI):		
*EZID	ez	zid.cdlib.org	

• "Long-term identifiers made easy"

EZIC) sy	Login Contact Us
Home Manage ID	s Create IDs Lookup ID Demo	
Why EZID? Understandir EZID (easy-eye-dee) ma		Logout Profile Contact UC3 Help Logged in as Stephen Abrams
• create identifiers for an	Met Collection home Change collection 🔻	
 store citation metadata update current URL loc use EZID's programmin choose from a variety o try out EZID anytime and 	Object: ark:/b7297/x2bv7dj2 Merrit > Collection: UC Berkeley Linguistics Dept California Language Archive > Object: ark:/b7297/x2bv7dj2	Object Lookup Go Search object author, title, date or identifier.
	Bloomfield, Leonard. [Eastern Ojibwa notebook #14] (undated)	
Latest news from EZID	permanent link: http://n2t.net/ark:/b7297/x2bv7dj2 primary identifier: ark:/b7297/x2bv7dj2 title: [Eastern Ojibwa notebook #14] creator: Bloomfield, Leonard	Current Version
	date: undated local id: Bloomfield.003.001	Version 2: 2013-06-29 06:22 PM UTC = Bloomfield.003.001.0001.tif = Bloomfield.003.001.0002.tif = Bloomfield.003.004.0002.tif
NE ZID	Iast modified: 2013-06-29 06:25 PM UTC created: 2013-06-29 06:21 PM UTC total size (all versions): 3.1 GB	 Eloomfield.003.001.0003.tif Eloomfield.003.001.0004.tif
<u><u></u>™⊏∠ID</u>	total versions: 2	53 more files
Long-term identifiers made easy 2014	Download object	Prior Versions
2014		Version 1: 2013-06-29 06:21 PM UTC

Merritt

"Preservation and access"



Merritt is a new cost-effective repository service from the University of California Curation Ce the UC community manage, archive, and share its valuable digital content. Use Merritt to pr preservation of digital assets, share your research with others or meet the data sharing and requirements of a grant-funded project.

Merritt is built upon a Micro-Services approach to digital curation based on devolving curation of independent, but interoperable, services that embody curation values and strategies. Sinc services is small and self-contained, they are collectively easier to develop, deploy, maintair This design approach is detailed in the UC3 Curation Foundations paper.

Learn more about Merritt:

- Handout: Merritt Uses and FAQ
- Merritt User Guide

Need a Merritt Account? Fill out the Contact UC3 form or email uc3@ucop.edu.

> Merritt is a service of the University of California Curation Cen Copyright © 2009-2014 The Regents of the Univ Privacy policy | Image credits

- No prescriptive requirements on content genre, type, format, structure, or metadata
- Strong versioning maintains complete change history
- Restricted or public access under your control
- Enforceable data use agreements (DUAs)
- Storage replication to UCLA and UCSD, with ongoing auditing
- Integration with EZID and DataONE



• Proactive preservation analysis, planning, and intervention



merritt.cdlib.org

DataONE

• "Data observation network for Earth"

				Connect	
ta Observation Network for E	Earth			V You Tube	∎F ŧ ୬å
About Participate	Resources	Education	Data		
	A DataONE Search	Tool for Scientific	Data		
Search For: sour	rce water		in and "moisture content")		Results/Page 10 SEARCH Only results with data.
now/Hide Advanced Options	Clear All Help				
elded Search	0	R 💌	Collection Date during Collection Date Publication Date Either mm/dd4	WW mr	n/dd/yyyy Help clea
eographic Search					
Onice States	North Atlantic Ocean	Algeria Libya Egypt Mali Niger Chad Nigeria DR Congo	ikey Iraq iran Afghanistan Saudi Arabia India	Trailand	List Areas in: USA ③ WORLD ③ Search Area: overlaps ④ encloses ③ North West East South





- Cyberinfrastructure
 - Distributed grid of member and coordinating nodes
 - Aggregated discovery
 - Investigator's toolkit

• Community



• "Data sharing made easy"

Library & Oash UNIVERSITY OF CALIFORNIA									
About	Search Data	Share Data	My Datasets						
Searc or Browse	h for Data Go all data	Share Dat Make your data cr Get credit for your Ensure reproduci Promote reuse Meet funder requi Learn more	ount r data bility						
Terms of Use FAQ	Contact Us								



datashare.ucsf.edu

- Preservation repositories are complex systems
- Far too often, their interfaces are complicated and meant only for IT professionals and archivists
- Dash provides a set of user-friendly screens to step through the process:
 - Select/upload files associated with a dataset
 - Augment with descriptive metadata
 - Review that the dataset meets requirements and is ready
 - Submit to the Merritt preservation repository with optionally replication to DataONE





datashare.ucsf.edu

• Upload dataset files

UNIVERSITY OF CALIFORNIA	Logged in	as Perry.Willett-ucop.edu@ucop.edu > Logout					
About	Search Dat	ta	Share Data	My Datasets			
STEP ONE: Prepare	STEP TWO: Describe	STEP THREE: Upload	STEP FOUR: Review & Submit				
Upload your datasets to Dash							
		here files	of the Dash Ensure that human subj properly de- Ensure data	a have read the terms Data Use Agreement all data relating to jects has been identified a files are the most up- complete versions			
Go Ba			 Dash accept Generally s you upload into one of the second se	ots all file formats. peaking, the data files into Dash should fall the following types: dsheet (csv, tab-			



datashare.ucsf.edu

• Add descriptive information

UNIVERSITY OF CALIFORNIA	NG MADE EASY	Logged in as <i>Perry.Willett-ucop.edu@ucop.edu</i> > Logout		
About	Search Data	S	ihare Data	My Datasets
	EP TWO: escribe	STEP THREE: Upload	STEP FOUR: Review & Submit	
Describe yo * Required Field	ur dataset			
* Dataset Title 💡	Wind in Cuerda del P	ozo. February 2012]	
* Institution 🥹	UC Berkeley	•		
* Data Type 💡	Presentation	•		
* Data Creator(s) 💡	Lastname, Firstname		+ Additional Creator	
	Aguilar, Fernando delete			
Keywords 🥥	Keyword		+ Additional Keyword	
	_	_		



datashare.ucsf.edu

• Review the dataset

UNIVERSITY OF CALIFORNIA	DATA SHARING MADE EASY			Logged in as Perry	y.Willett-ucop.edu@ucop.edu
About	Search Da	Ita	Share Data		My Datasets
STEP ONE: Prepare	STEP TWO: Describe	STEP THREE: Upload	STEP FOUR Review 8	≳ & Submit	
Review Metadata	before submit	ting to D	ash		
Dataset Title:	Wind in Cuerda del Pozo. Februa	гу 2012			
Institution:	University of California, Berkeley				
Publication Year:	2014				
Data Type:	Presentation				
Data Creator:	Aguilar, Fernando				
Keywords:	Spain Wind				
Citation(s):	 Show Citations 				
Abstract:	 Show Abstract 				
Methods:	Show Methods				



datashare.ucsf.edu

• Submit to a repository

F	Record contains all recommended fields		
	Go Back Edit Description		
L. L.	Jploaded Files		
F	Record must include at least one file		
	Upload More Files		
F	Ready to go?		
	Submit to Dash		
Terms of Use	FAQ Contact Us		
Calments of Calments	Dash is a service of the University of California Digital Library. Copyright © 2014 The Regents of the University of California	Dash is a collaborative project between the UC Libraries and the UC Curation Center.	*UC 3



datashare.ucsf.edu

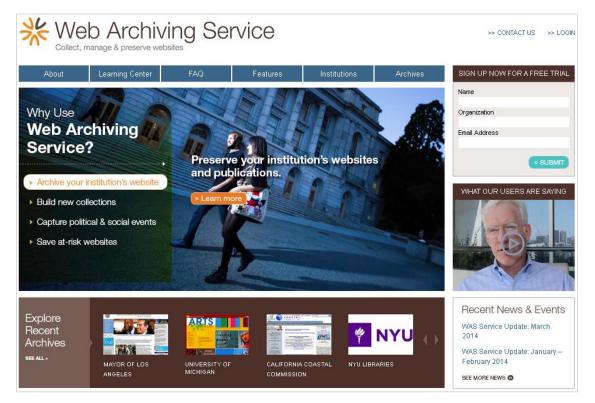
• Search/browse and discovery

About	Search Data	Share Data	My Datasets
Select a Data	iset	3	Search datasets Go!
Campus			
• UC Berkeley (8)	Sorted by: title	▼ Go!	
UC Los Angeles (2)			
UC San Francisco (2)	Phylogenomic inferen	nce of protein molecular functi	ion: supplementary data
Author	by Sölander, Kimmer at UC Berkeley	I	
Willett, Perry (2) Aguilar, Fernando (1)	 Show abstract 		
Banerjee, Utpal (1)			
Beylkin, Gregory (1)	DMPTool usage		
Bloom, Joshua S. (1)	by Willett, Perry Wil at UC Berkeley	lett, Perry	
Butler, Nathaniel R. (1)	Show abstract		
more		efunction as an unconstrained	sum of Slater
Keyword	determinants: datase	t	
Amino acid sensing (1	by Beylkin, Gregory at UC Berkelev	Mohlenkamp, Martin J. Pérez, Ferr	nando



datashare.ucsf.edu

• *"Capture and preserve the web"*





was.cdlib.org webarchives.cdlib.org

• The web is a volatile environment

	berkeley.org	_ □ <mark>×</mark> @☆ <mark>》</mark> =
occupyberkeley.org		Search Ads
Related Links Car Insurance Brokers Cheapest Car Insurance	Sponsored Listings GEICO Car Insurance www.geico.com \$100? \$200? \$300? How much could you save with GEICO? Get a quote.	•
Quotes Best Car Insurance Companies	\$19 Car Insurance - New auto-insurance.quotelab.com/MI Get a Free Quote & Save 55-75%! (Michigan Residents Only)	Ð
Cheap Car Insurance Rates	AAA Home Insurance www.aaa.com Get Quality Home Insurance at Great Rates. Call For a Free Quote.	₽
Young Driver Car Insurance Military Auto Insurance	Restaurant Insurance cransoninsuranceagency.com Save up to 25% on your Restaurant Insurance	•
Car Insurance in Washington DC State Auto Insurance	"Cheapest Auto Insurance" www.onlineautoinsurance.com Instant Quotes From Up To 30 Co's Save Time and Money! Takes ~4 min.	Ð
Company Cheapest Insurance Company Auto Insurance Quote In		
🕂 🚞 🗴 🔤 🚺 📴		▲ 📭 📆II (4)) 5/26/2014



was.cdlib.org webarchives.cdlib.org

• WAS captures and preserves important web content



was.cdlib.org

webarchives.cdlib.org

BITSS Summer Institute

💥 Web Archiving Service

Collect, manage & preserve websites

• WAS captures the web over time

Captured Od/29/12 08:16 AM Occupy Web Archive O5/02/12 11:32 AM UCLA Library O5/20/12 08:09 AM Home About Site List Search Het O6/04/12 06:04 AM Occupy Berkeley O7/02/12 06:42 PM Description Occupy Berkeley O7/02/12 06:32 PM Description Occupy Berkeley O1/02/12 06:32 PM Description Occupy Berkeley O1/02/13 06:33 PM O/02/12 01:6 AM O2/02/13 06:33 PM O/02/13 06:32 PM O1/02/13 06:32 PM O/02/13 06:32 PM O3/02/13 06:32 PM O/02/13 06:32 PM O3/02/13 06:32 PM O/02/13 06:32 PM O3/02/13 06:32 PM Of/02/13 06:32 PM O5/02/13 06:33 PM Of/02/13 06:32 PM O5/02/13 06:33 PM Of/02/13 06:32 PM O5/02/13 06:33 PM Of/02/13 06:32 PM Of/02/13 06:32 PM Of/02/13 06:32 PM Of/02/13 06:33 PM Of/02/13 06:33 PM Of/02/13 06:33 PM Of/02/13 06:33 PM <thof 02<="" th=""><th>← → C ↑ bwebarchives.cdlib.org/site/sw1dj59k3r</th><th></th><th>🖾 🖓 🔊</th></thof>	← → C ↑ bwebarchives.cdlib.org/site/sw1dj59k3r		🖾 🖓 🔊
	Vesterday's web; today	04/29/12 08:16 AM 05/02/12 11:32 AM 05/07/12 02:33 PM 05/13/12 06:59 PM 05/20/12 08:09 AM 06/04/12 06:04 AM 07/02/12 06:42 PM 08/02/12 06:41 PM 10/04/12 04:58 PM 11/02/12 06:32 PM 12/05/12 04:00 AM 01/02/13 06:33 PM 02/02/13 06:35 PM 03/02/13 06:32 PM 04/02/13 06:32 PM 05/02/13 06:33 PM	
	듹 🚞 XII 🔤 💽 📴		▲ 🏴 📆il 🕪 1:32 PM 5/26/201
	Wab Arabiving Sanias	was.cdlib.org	
was.cdlib.org	Collect, manage & preserve websites	webarchives edlib ora	

webarchives.cdlib.org

- WAS provides curators with tools to capture the free web:
 - Schedule web crawls on regular or customized basis
 - Focus on website itself or include linked sites
 - Brief 1-hour or full 36-hour crawls
 - Analyze results with a range of reports
 - Search across captured websites
 - Keep archive restricted, or provide public access
 - Fee-based service



was.cdlib.org webarchives.cdlib.org

- WAS includes archives based on events:
 - 2003 California recall election
 - 2007 Southern California wildfires
- Thematic archives:
 - Grateful Dead archives
 - US Labor unions and organizations
 - California political blogs
- Comprehensive archives of web-domains:
 - Emory University
 - University of Michigan

Web Archiving Service

was.cdlib.org webarchives.cdlib.org

Service takeaways

- DMPTool Create data management plans required by funders or journals using campus resources
- DataUp *Curation services tailored for tabular datasets*
- Dash Simplified interfaces for repository submission
 and discovery
- EZID / Core infrastructural services generally hidden Merritt beneath simple intuitive interfaces
- WAS Curation services tailored for web-published content and data

tion Off Reprinters Party	terry wave two furtherna A	ert* Oues
Data Manage Planning Tool Create, review, and share	1 House	
data management plans to institutional and funder reg		
inettutional and funder req		Ct DAPTOOL HELP
	urements	

Despheren	Terring Selling Connect Series Series	may being t
VILLE		And data because of
DataUp		MUTH MAN A ADDRE. Marginal Charles
Contraction of the second	· Aritim Albin	many dependences of the support of the data
- Oneck for Best Practice - Orests Microbia		and this many our wat
- GetOvet	Printing on Dentity of Least resident Spectrum	section () and () and ()
Armestres		Conception Conception
		manufactures.
Argoni & Dave B		And your world to be

Search for Data	Share Data - Marc Dat Data - Marc Dat Data - Marc Data	
	de:	



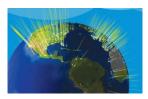
Summary

- Good data management practice is critical to the success of the academic enterprise and scholarly advancement
- Management solutions should be integrated into existing research systems and workflows
- The UC Libraries are a natural partner for data management advice and solutions
- UC3 offers a comprehensive roster of innovative and intuitive curation services applicable across the data and scholarly lifecycle









For more information

- UC Curation Center www.cdlib.org/uc3 datapub.cdlib.org uc3@ucop.edu
- DMPTool dmptool.org
- DataUp/ONEShare dataup.org
- Dash
 - EZID
 - Merritt
 - DataONE
- WAS

- datashare.ucsf.edu
- ezid.cdlib.org
- merritt.cdlib.org
- dataone.org
- was.cdlib.org





Build your Data Management Plan

Contraction Describe, manage & share your data









