Washington University in St. Louis Washington University Open Scholarship

IASSIST & DCN - Data Curation Workshop

Workshop Schedule

Dec 12th, 9:00 AM - 10:00 AM

Transform Presentation

Lisa Johnston University of Minnesota - Twin Cities, ljohnsto@umn.edu

Jennifer Moore *Washington University in St. Louis,* j.moore@wustl.edu

Follow this and additional works at: https://openscholarship.wustl.edu/data-curation-workshop-2017

Part of the Library and Information Science Commons

Johnston, Lisa and Moore, Jennifer, "Transform Presentation" (2017). *IASSIST & DCN - Data Curation Workshop*. 7. https://openscholarship.wustl.edu/data-curation-workshop-2017/schedule/Schedule/7

This Presentation is brought to you for free and open access by the Conferences and Symposia at Washington University Open Scholarship. It has been accepted for inclusion in IASSIST & DCN - Data Curation Workshop by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

Welcome Back

Day 2 of the Data Curation Workshop

Hashtag #DCW2017

$\begin{array}{c} \textbf{C} \twoheadrightarrow \textbf{U} \twoheadrightarrow \textbf{R} \twoheadrightarrow \textbf{A} \twoheadrightarrow \textbf{T} \twoheadrightarrow \textbf{E} \\ \downarrow \\ \end{array}$ Check files and read documentation

$C \Rightarrow U \Rightarrow R \Rightarrow A \Rightarrow T \Rightarrow E$ $\downarrow \downarrow$ Understand the data

$C \twoheadrightarrow U \twoheadrightarrow \mathbf{R} \twoheadrightarrow A \twoheadrightarrow T \twoheadrightarrow E$ $\downarrow \downarrow$ Request missing information

$C \twoheadrightarrow U \twoheadrightarrow R \twoheadrightarrow A \Rightarrow T \Rightarrow E$

Augment metadata for findability

Minute Paper Questions

- What do you do if the depositor has not replied to any of your emails?
- How much time do you spend curating a dataset?
- How to apply "science-y curation" to humanities work?
- Support
 - How do you justify staff funding to your administration?
 - Advice on generating support from administrators and researchers?
- How do you set the scope of your repository, 47 activities is too many?
- What is your staffing model and implementation plan for the DCN?
- How do you get more data sets to curate?
- Security procedures? Do you put SIPs in quarantine?

$C \Rightarrow U \Rightarrow R \Rightarrow A \Rightarrow T \Rightarrow E$ $\downarrow \downarrow$ Transform file formats

Defined

Transformation is the process of converting data from one format (e.g. a database file, XML document, or Excel sheet) to another.



Reasons/Benefits of Transforming

Brainstorm why to transform:

- Consider how it benefits different stakeholders (who are they)
- Consider different scenarios where the transformed data creates advantages

Reasons/Benefits of Transforming

- Users that do not have native software
- Future migrations
- Common formats that many people have/can access
- Storage efficiency
- Marketing for what the dataset contains (preview)
- Obsolescence
- Be careful not to lose information
- Accessibility

File Format Transformations

Brainstorm with your group:

- What are some format transformations for your datasets? Discuss what that means for different data types

 - What are the challenges?

Example File Format Transformations

- 1. CZI (microscope image) native software exports as TIFF, JPEG, FITS (astro image file)
 - a. WikiData tracks software and file formats for preservation
 - b. Omero, Bioformats are tools that help
- 2. XLS \Rightarrow CSV (what about formulas??)
- 3. Chemdraw \Rightarrow JPG, 001, .opj, .tri \Rightarrow ??
- MP4 ⇒ adding CC (good practice) keep both, web archiving ⇒ screenshot, IA, link to live site
- 5. .shp (geocoded xls) ⇒ retain (useful info) ⇒ csv (tabular), extract metadata
 a. FME tool for conversion but ArcGIS too
- 6. CSV, PDF \Rightarrow good. (other ex, QuarkExpress inDesign)

Preservation File Formats for Long-term Access



Text	MS Word	PDF, TXT, HTML
Images	Photoshop	TIFF
Video/Media	Quicktime	MPEG4
Database	MS Access	DBF
Tabular Data	MS Excel	CSV
Presentations	MS Powerpoint	PDF (unencrypted)
Sound/Music	Windows Media	WAV (uncompressed)

*Be conscious about the risks of compressing your files or migrating to a file format that has different affordances than the original. See more at <u>http://guides.library.cornell.edu/ecommons/formats</u>

Excel Archival Tool

- Automated conversation process for
- Microsoft Excel → CSVs but also captures
 - Charts and figs as PNGs
 - Formulas
 - Cell formatting and style
- Generates a report on the archival outputs



Case Study: MS Excel



Download from GitHub: http://z.umn.edu/exceltool

Case Study: GIS

Point Dataset:

Some locations generated using **batch geocode** of addresses using **ArcMap** in WGS 1984

Some locations generated using **point-by-point** address selection in **AGOL** in WGS 1984 Web Mercator Auxiliary Sphere

Case Study: GIS

1. Open dataset from AGOL to desktop	6. Value standardization
2. Export copy to .shp.	7. Export to table
3. Projection	8. Save as CSV
4. Calculate geometry	
5. Remove columns	

Calculate the geometry to make the Latitude and Longitude appear in the table.

Using datasets coordinate system presents an issue

		Side	City	Region	RegionAbbr	Country	X	Y	Xmin	Xma
		L	East St. Louis	Illinois	L	USA	-90.139155	38.633856	-90.140155	-90.13
							-90.135891	38.636169	0	
		R	West Alton	Missouri	MO	USA	-90.130898	38.839825	-90.141898	-90.11
		L	East St. Louis	Illinois	L	USA	-90.101083	38.630513	-90.102083	-90.100
		R	Wood River	Illinois	L	USA	-90.097818	38.860969	-90.098818	-90.096
		R	Wood River	Illinois	L	USA	-90.097818	38.860969	-90.098818	-90.096
		R	East St Louis	Illinois	L	USA	-90.096381	38.609487	-90.097381	-90.09
		R	Belleville	Illinois	L	USA	-90.0918	38.561664	-90.0928	-90.0
							-90.078868	38.636904	0	
							-90.067966	38.66155	0	
		R	Caseyville	Illinois	L	USA -90.0485		38.612063	-90.049586	-90.04
		L	Wood River	Illinois	L	USA	-90.043178	38.872519	-90.046178	-90.040
		R	Belleville	Illinois	L	USA	-90.029785	38.541592	-90.030785	-90.021
		R	Belleville	Illinois	L	USA	-90.011116	38.547491	-90.012116	-90.010
		R	Fairview Heights	Illinois	L	USA	-90.003541	38.579752	-90.004541	-90.002
							-89.997214	38.718609	0	
			Edwardsville	Illinois	L	USA	-89.996098	38.793121	-89.997098	-89.99
			Edwardsville	Illinois	L	USA	-89.996098	38,793121	-89.997098	-89.99
		R	Belleville	Illinois	L	USA	-89.988304	38,511691	-89.989304	-89.98
		R	Fairview Heights	Illinois	L	USA	-89.987945	38.585576	-89.988945	-89.986
		L	Collinsville	Illinois	L	USA	-89.987669	38 66957	-89.988669	-89.986
		R	Belleville	Illinois	1	USA	-89 987251	38 512621	-89 988251	-89.98
		P	Belleville	Illinois		USA	_89 985873	38 513623	-89 986873	-89 98
		D	Belleville	Illinois	1	USA	-89 985832	38 513623	-89 986832	-89.98
		1	Belleville	Illionie	1	USA	-00.000002	38 510551	-89 985287	-00.00
		1	Belleville	Illinois		USA	-03.304207	20 510331	-09.903207	-03.30
		1	Belleville	Illinois		USA	-03.304243	20 510074	-09.903243	-03.30
		L	Delleville	IIIIIIOIS	IL	USA	-03.304243	30.510074	-03.303243	-03.30
		0	Dalla illa	We also		110.4	-09.904040	38.500642	0	00.001
		ĸ	Belleville	IIIInois	IL I	USA	-89.983267	38.512597	-89.984267	-69.98
		L	Belleville	IIIInois	IL.	USA	-89.976796	38.5111/1	-89.977796	-89.97
		R	Belleville	IIIInois	IL.	USA	-89.976007	38.510616	-89.977007	-89.97
	_	R	Belleville	Illinois	L	USA	-89.976007	38.510616	-89.977007	-89.97
	Calculate Ge	ometry				×	-89.976007	38.510616	-89.977007	-89.97
							-89.976007	38.510616	-89.977007	-89.97
	Property	X	Coordinate of Point				-89.973051	38.508182	-89.974051	-89.972
	Property.	^	Cool dinate of Point				-89.970797	38.503698	-89.973797	-89.96
	Coordinate	System					-89.912271	38.594339	0	
	Ollse coo	rdinate s	vstem of the data sour	ce:			-89.894711	38.725431	-89.896711	-89.892
	DCC N	NCC 100	A Mah Masantas Augilia	cu Calence			0	0	0	
	PCS: 1	NG2 190	4 Web Mercator Auxilia	ry sphere				0	0	
	0.000	adapte a	where a false data from				0	0	0	
	Use coo	roinate s	ystem of the data fran	ie:			0	0	0	
	GCS: 1	WGS 198	4				0	0	0	
							0	0	0	
							0	0	0	
	Units:	M	eters [m]			~	0	0	0	
							0	0	0	
	Calculate	selected	records only				0	0	0	
	About calculating geometry OK Cancel				0	0	0			
				ancel	0	0	0			
	-						0	0	0	
			St. Louis	Missouri	MO	USA	0	0	0	
			Clayton	Missouri	MO	USA	0	0	0	
		-	Edwardsville	Illinois	L	USA	0	0	0	
			St Louis	Missouri	MO	USA	0	0	0	
		-	St Louis	Missouri	MO	USA	0	0	0	
			St Louis	Missouri	MO	USA	0	0	0	
		1	St. 20013	maaouli	1			0	0	

X is calculating in Web Mercator Auxiliary Sphere, which is projected. Better to keep it WGS 1984, as Y.

Clear that dataset needs to be projected in same coordinate system.

City	Region	RegionAbbr	Country	X	Y
St. Louis	Missouri	MO	USA	-10885069.6	0
Chesterfield	Missouri	MO	USA	-10081418.6	38.649478
Chesterfield	Missouri	MO	USA	-10078055.9	38.673915
Manchester	Missouri	MO	USA	-10075832.1	38.592948
Saint Louis	Missouri	MO	USA	-10074749.5	38.649374
Valley Park	Missouri	MO	USA	-10073719.8	38.549873
Wildwood	Missouri	MO	USA	-10073083.2	38.596327
Maryland Heights	Missouri	MO	USA	-10070929.3	38.746372
St. Louis	Missouri	MO	USA	-10069806.4	0
Fenton	Missouri	MO	USA	-10069611.7	38.542593
Des Peres	Missouri	MO	USA	-10068424.9	0
Saint Louis	Missouri	MO	USA	-10068345.0	38.701745
St. Louis	Misouri	MO	USA	-10068134.2	0
Maryland Heights	Missouri	MO	USA	-10067976.8	38.705587
St. Louis	Missouri	MO	USA	-10067888.5	0
Maryland Heights	Missouri	MO	USA	-10067829.0	38.736082
Bridgeton	Missouri	MO	USA	-10067431.4	38.756542
Saint Louis	Missouri	MO	USA	-10067210.2	38.667188
Sunset Hills	Missouri	MO	USA	-10066800.2	38.545277
Saint Louis	Missouri	MO	USA	-10065845.6	38.602997
Saint Louis	Missouri	MO	USA	-10064292.2	38.555388
Saint Louis	Missouri	MO	USA	-10064199.2	38.684463
Saint Louis	Missouri	MO	USA	-10064079.1	38.578898
Saint Louis	Missouri	MO	USA	-10064038.5	38.630724
St.Louis	Missouri	MO	USA	-10064008.9	0
Kirkwood	Missouri	MO	USA	-10063992.0	38.580664
St Ann	Missouri	MO	USA	-10063964.3	38 726449

GIS Discussion

Example was on a point dataset, but what if you had:

-polygon data

-line data

OR

-raster data

Case Study: Video Transcription

ELAN software

Open source download (<u>https://tla.mpi.nl/tools/tla-to</u> <u>ols/elan/</u>)

Native files .eaf (mac) or .pfsx (Win)

Runs mov and avi files and allows for transcription

Outputs include XML



- 0 >

Brainstorm at your tables

Challenges for implementing data curation at your institution