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

Welcome Back

Day 2 of the Data Curation Workshop

Hashtag #DCW2017

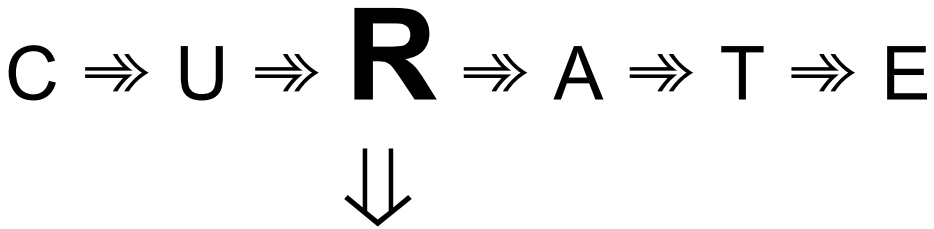
C ⇛ U ⇛ R ⇛ A ⇛ T ⇛ E

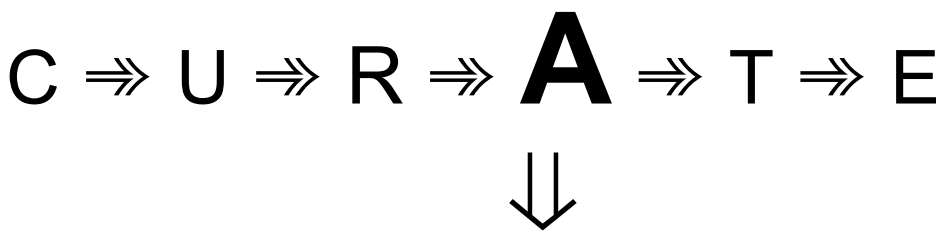
⇓

Check files and read documentation

C ⇛ **U** ⇛ R ⇛ A ⇛ T ⇛ E

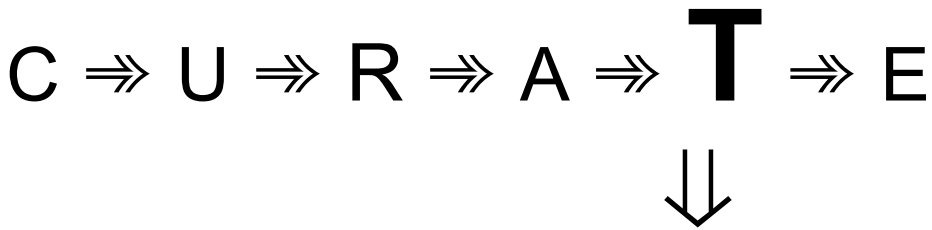
Understand the data


C ⇛ U ⇛ R ⇛ A ⇛ T ⇛ E
⇓
Request missing information


C ⇛ U ⇛ R ⇛ A ⇛ T ⇛ E
⇓
Augment metadata for findability

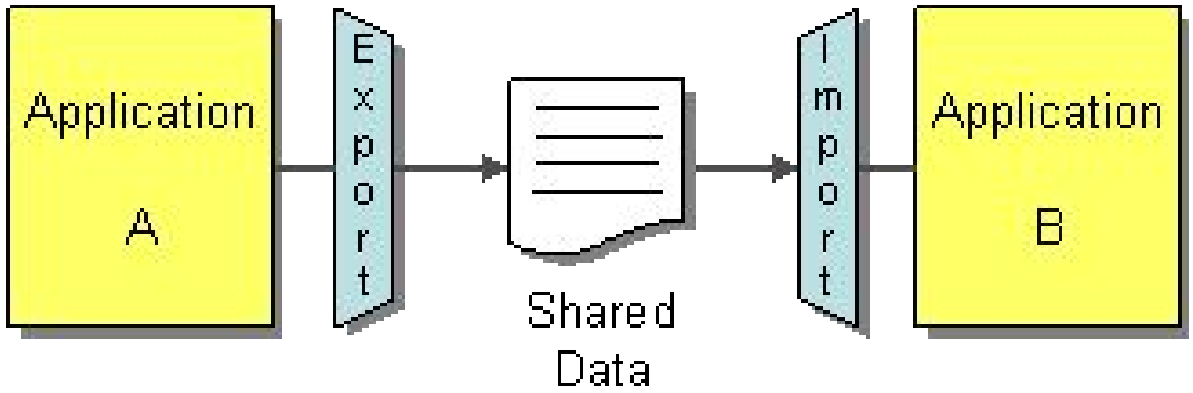
Minute Paper Questions

- 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 ⇛ U ⇛ R ⇛ A ⇛ T ⇛ E
⇓
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 ⇒ CSV (what about formulas??)
3. Chemdraw ⇒ JPG, 001, .opj, .tri ⇒ ??
4. 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 ⇒ good. (other ex, QuarkExpress inDesign)

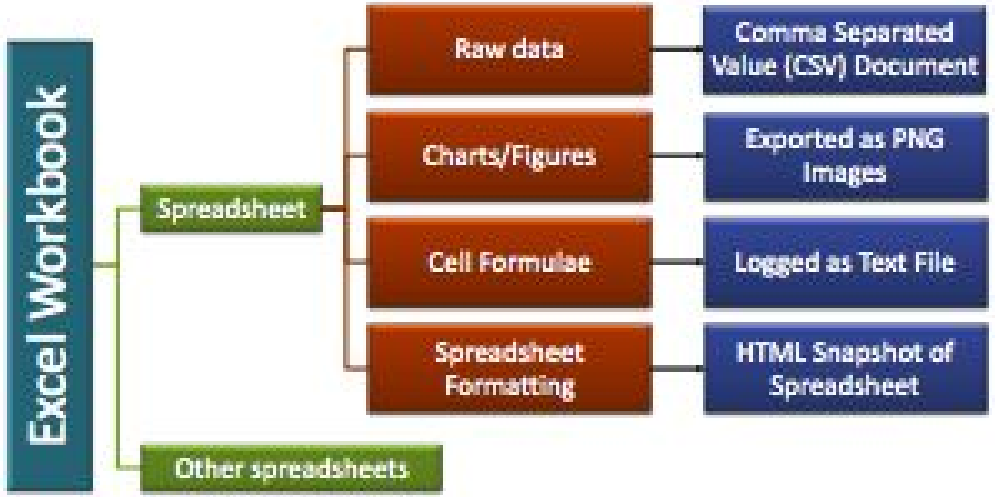
Preservation File Formats for Long-term Access

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 http://guides.library.cornell.edu/ecommons/formats

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



Citation: McGrory, John. (2015). Poster for "Excel Archival Tool: Automating the Spreadsheet Conversion Process". Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/171966.

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

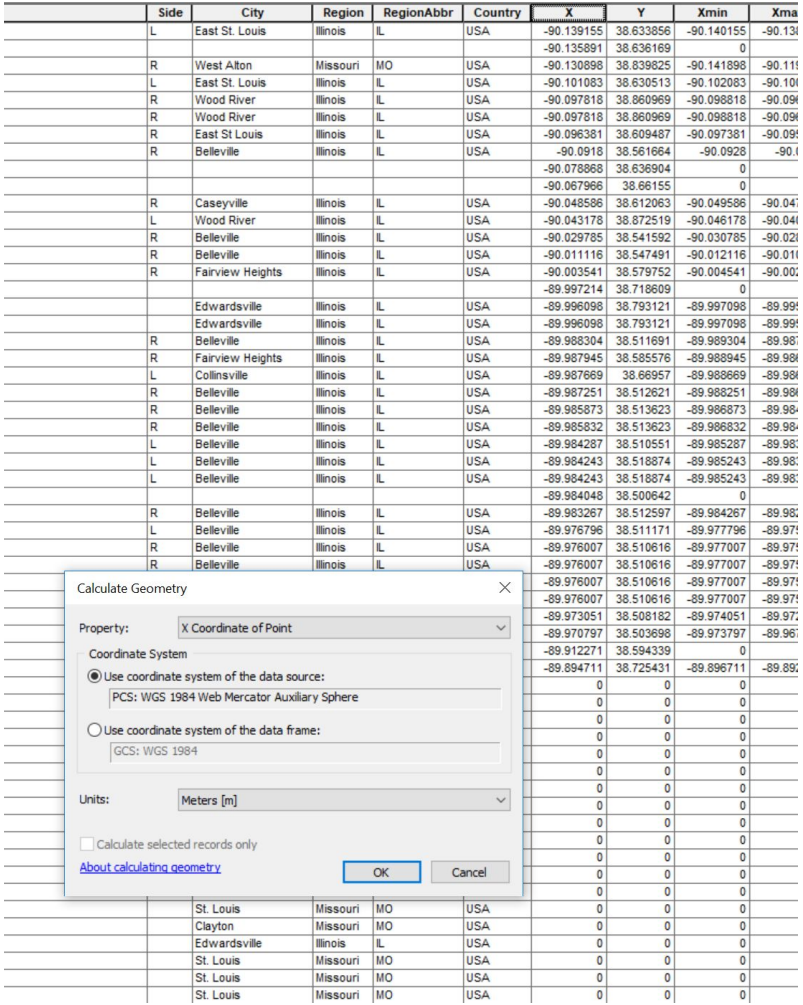
Some locations generated using **point-by-point** address selection in **AGOL in** WGS 1984 Web Mercator Auxiliary Sphere

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



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

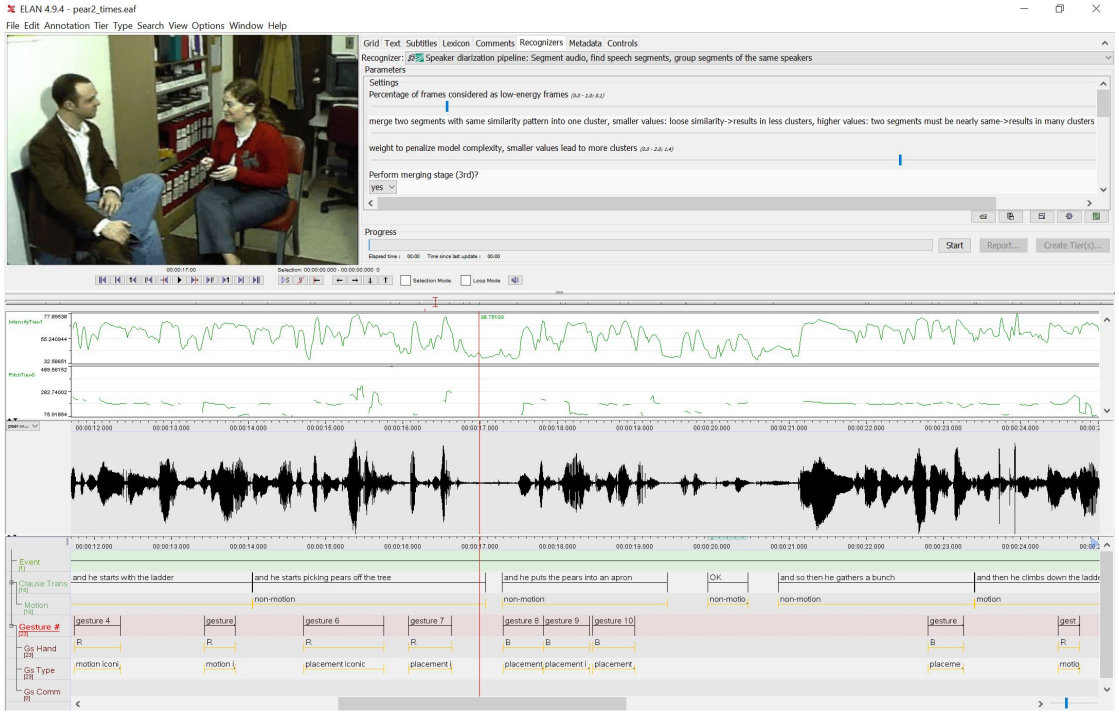
ELAN software

Open source download (https://tla.mpi.nl/tools/tla-tools/elan/)

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

Runs mov and avi files and allows for transcription

Outputs include XML



Brainstorm at your tables

Challenges for implementing data curation at your institution