DCN Planetears

Geospatial Data Curation:
an introduction
Module: Transformations

Module Objectives: Transformations

This module has three objectives. At the end of the module, learners should be able to:

1. Understand common reasons for recommending transforming file formats
2. Assess benefits and downsides of common file formats
3. Perform transformation of vector data in QGIS

The module has Lecture and Activity components.

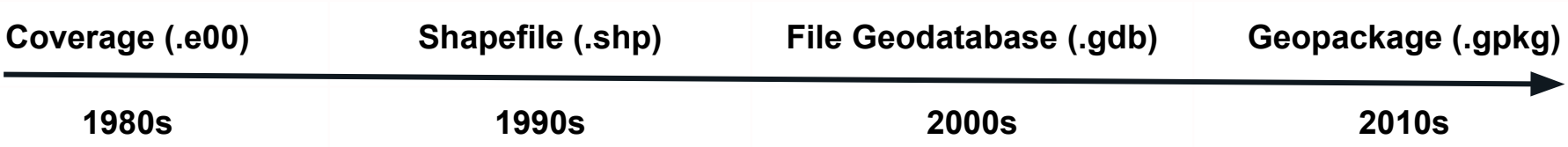
Reason That It Is Hard

There isn’t a clear archival standard file format for geospatial data.

Vector Geospatial File Formats Over Time

CURATED

Geospatial file formats have changed a lot over the past 40 years and they continue to change



(.SHP and .GDB are most common even though there are reasons they might not be the best pick)

Proprietary vs Open File Formats

Some of the most common and interoperable geospatial file formats are proprietary.

Many proprietary geospatial file formats can be opened with open source software.

Questions to Consider

C
U
R
A
T
E
D

Transformation:

1. How complicated would it be to represent the file in an open format?
2. How much effort would it take to make that transformation?
3. How much more accessible are you actually making that data through transformation?

Library of Congress Recommended Geospatial Formats

CURATED

Preferred formats include:

- Shapefile
- GeoTIFF
- Esri File Geodatabase
- OGC GeoPackage
- GeoJSON

“Most complete data (all layers, appendices), **even if proprietary,** with a preference for preserving the native format and projection of the data”

Why might you want to transform file formats?

- To simplify the file format
- To increase the number of software that can open a file
- To provide access to a data held in proprietary data formats that cannot be opened by open source tools

What should you watch out for when transforming files?

- The limitations and quirks of shapefiles
- Embedded metadata

Watch Out For: Shapefiles

CURATED

Shapefiles have lots of limitations and quirks!

Original:
Geodatabase (feature class)

date_sampling_test
12/27/2021 11:13:55 AM
12/27/2021 11:13:13 AM
12/13/2021
12/7/2021 6:00:00 AM
7/15/2019 5:00:00 PM
4/29/1985

Transformed:
Shapefile

date_sampl
12/27/2021
12/27/2021
12/13/2021
12/7/2021
7/15/2019
4/29/1985

Image credit: Melinda Kernik

Watch Out For: Researcher Transformed Shapefiles

CURATED

Watch out for situations where the researcher has already transformed the file format

Signs: The field names have been truncated and don’t match the documentation.

Actions: Make sure the researcher is aware of limitations for shapefiles

Watch Out For: Metadata

It’s very easy to lose metadata during transformation!

Especially:

- Transformations done using QGIS
- Transformations into open formats (like GeoPackage) using proprietary software

Actions: Check whether metadata is still embedded after transformation and provide alternative access if needed

Watch Out For: Additional Components in Project Files

CURATED

Geodatabases and Project files (Esri)

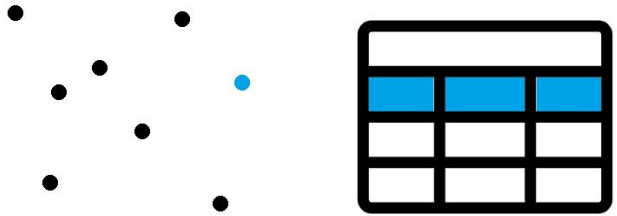
- can include toolboxes / code that require the proprietary software to work
- can have multiple-levels of embedded metadata that is very difficult to extract

Possible Transformations to Consider

CURATED

- Shapefile (Point data) → CSV
- Geodatabase →
  - GeoTiff (for rasters)
  - GeoPackage

Shapefile (Point data) → CSV

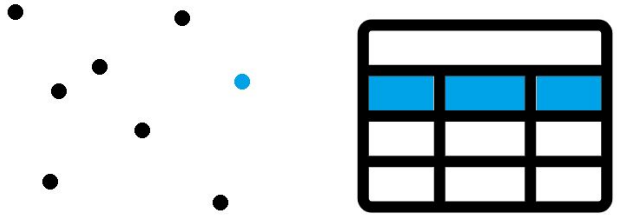


towns.dbf	spreadsheet
towns.shp	Geometry and index
towns.shx	
towns.sbx	

Possible Transformations to Consider: Shapefile Point Data to CSV: 02

CURATED

Shapefile → CSV



towns.dbf	spreadsheet
towns.shp	Geometry and index
towns.shx	
towns.sbx	

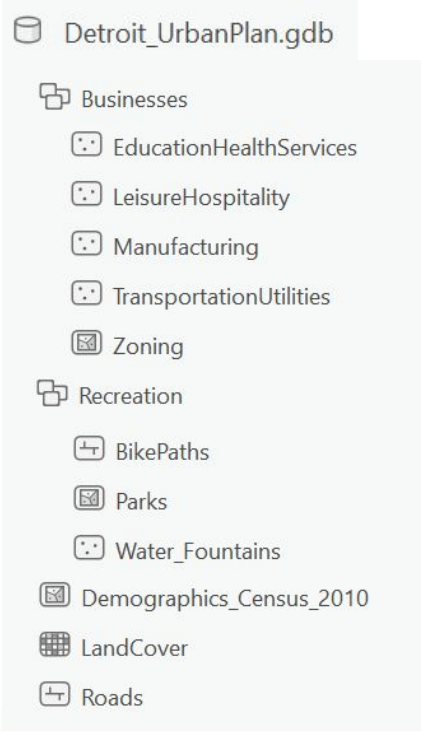
Latitude	Longitude
43.24	-95.61
46.13	-98.28
44.06	-92.74

Be sure that the Coordinate Reference System is noted in the metadata

Possible Transformations to Consider: Geodatabase: 01

CURATED

Geodatabase → ?



Possible Transformations to Consider: Geodatabase: 02

CURATED

Geodatabase → Shapefiles?

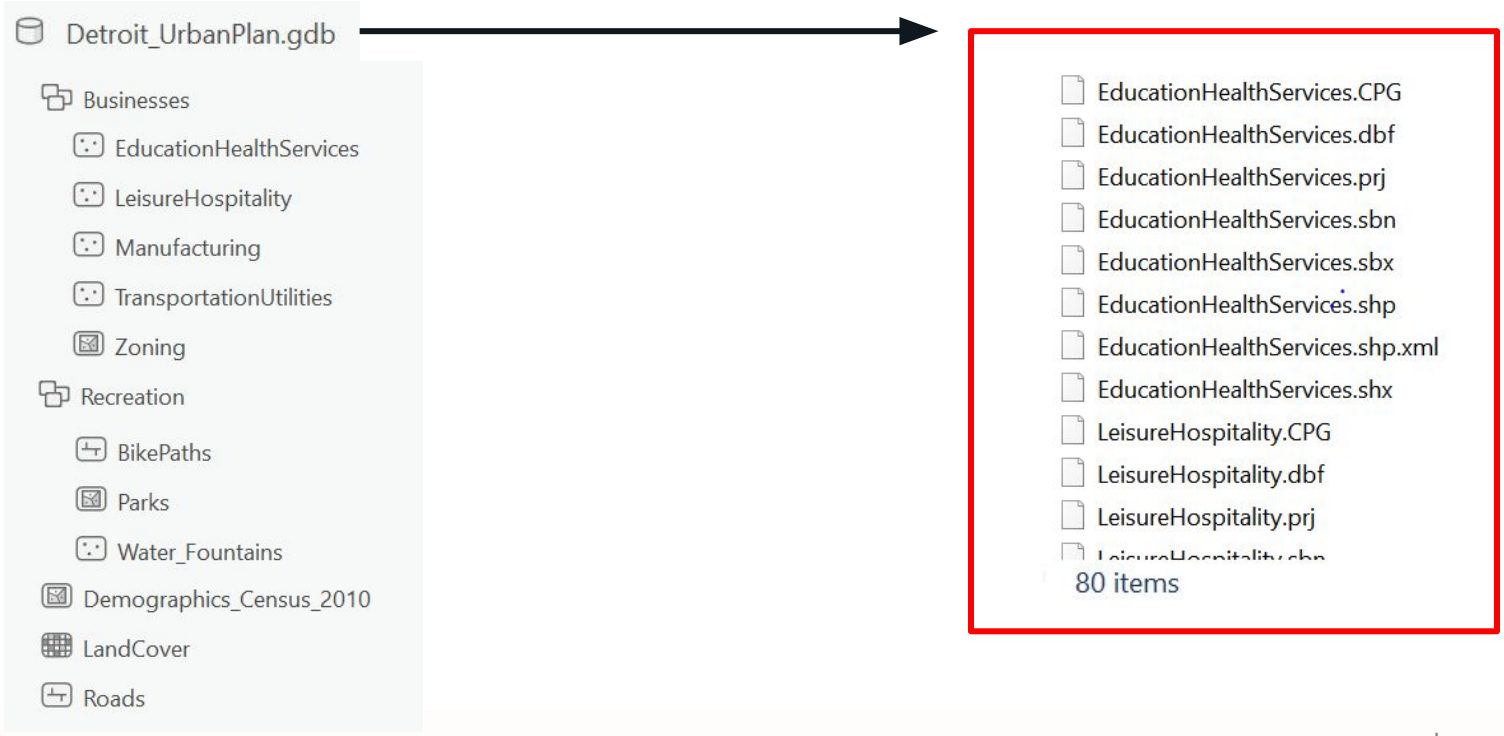
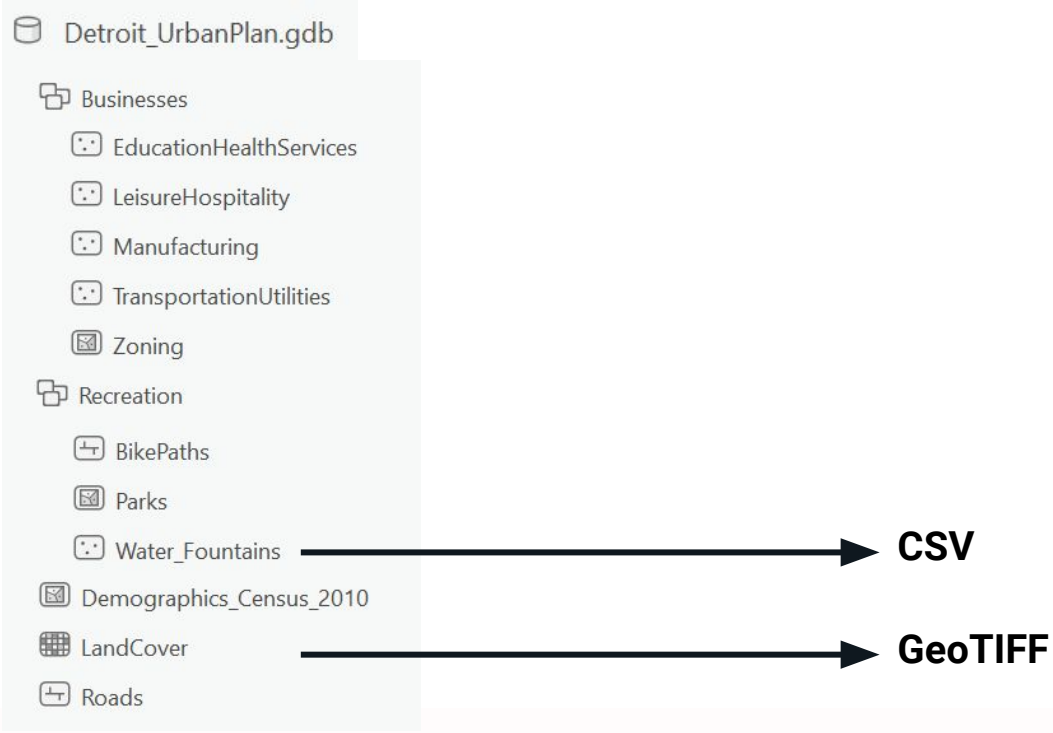


Image credit: Melinda Kernik

Possible Transformations to Consider: Geodatabase: 03

CURATED

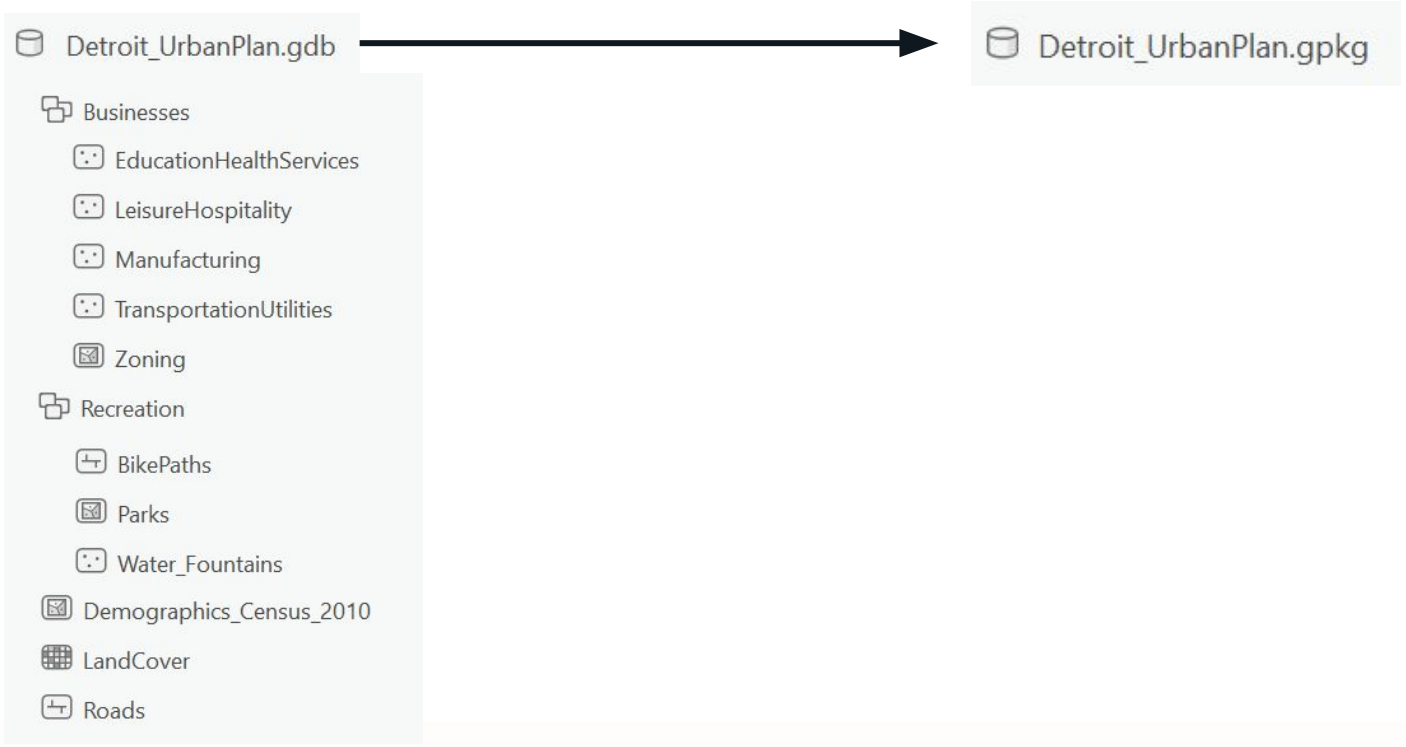
Geodatabase → Select layers?



Possible Transformations to Consider: Geodatabase: 04

CURATED

(Complex) Geodatabase → GeoPackage?



Discussion in Groups

Download and unzip (if needed) the example GIS dataset:
5_GIS_Transformations_Datasets.zip
available at https://bit.ly/5_GIS_Transformations

Review the file inventory “fileList_GIS_Transformations_Datasets.txt” What (if any) transformations would you suggest for this dataset? What questions would you have for the researcher to help you to decide what recommendations to make?

Discussion as a small group (5 minutes) and with the larger group (3 minutes)

Practice Transformations

Based on the group discussion, let’s practice transforming “5_GIS_Transformations_Datasets.gdb”!

Final Thoughts on Transformation

CURATED

- Weigh the amount of effort it would take to transform files against improvement to accessibility
- For complicated datasets, it is unlikely that transformation from proprietary into open formats will be possible without losing some essential pieces
  - Prioritize elements that will be inaccessible outside of proprietary software
  - Educate researcher about open formats for use in future projects
- If files are transformed, include steps taken in the documentation