Exploring Research Data Repositories with geoextent

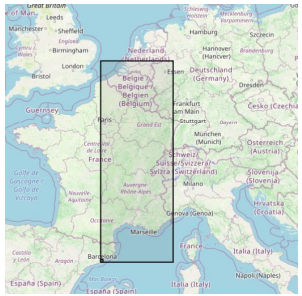
Sebastian Garzón, Daniel Nüst

Earth Cube Annual Meeting, 2021

June 15, 2021

geoextent

Python library for extracting **geospatial** and **temporal extent** of files and directories with multiple data formats



geospatial
bounding box
EPSG:4326

temporal
Period / time span
ISO8601 (Date)

Installing with `pip`*

```
pip install geoextent
```

Installing from source code

```
git clone
https://github.com/o2r-project/geoextent
cd geoextent
pip install -r requirements.txt
pip install -e .
```

*Build on top of

GDAL pandas NumPy

Supported formats:
GeoJSON (.geojson), **GeoTIFF** (.geotiff, .tif), **NetCDF** (.nc), **Shapefile** (.shp), **GeoPackage** (.gpkg), **GPS Exchange Format** (.gpx), **Geography Markup Language** (.gml), **Keyhole Markup Language** (.kml), **Tabular data** (.csv).

geoextent usage:

```
import geoextent.lib.extent as geoextent
```

```
# Individual files
> local_filepath = "../tests/testdata/shapefile/ifgi_denkpause.shp"
> geoextent_file = geoextent.fromFile( filepath = local_filepath,bbox = True, tbox = True)
....
bbox: [7.59, 51.96, 7.59, 51.96]  CRS: 4326 EPSG  tbox: ['2021-01-01', '2021-01-01']
```

```
# Folders / ZIP files
> local_dir_path = "../tests/testdata/folders/folder_two_files"
> geoextent_directory = geoextent.fromDirectory( path = local_dir_path, bbox = True,  tbox = True,
details = True)
....
[Combined]  bbox: [2.05, 41.31, 7.64, 51.97] CRS: 4326 EPSG  tbox: ['2018-11-14', '2019-09-11']
Individual (muenster_ring_zeit.geojson) : bbox: [7.60, 51.94, 7.64, 51.97] CRS: 4326 EPSG  tbox: ['2018-11-14','2018-11-14'']
```

```
# Data repositories
url_doi = "https://doi.org/10.5281/zenodo.820562"
> geoextent_url_doi = geoextent.from_repository(url_doi, bbox = True)
....
[Combined] bbox: [96.21, 25.55, 96.35,25.63] CRS: 4326 EPSG
```

Ind. File

Multiple Files

Local

Remote

zenodo

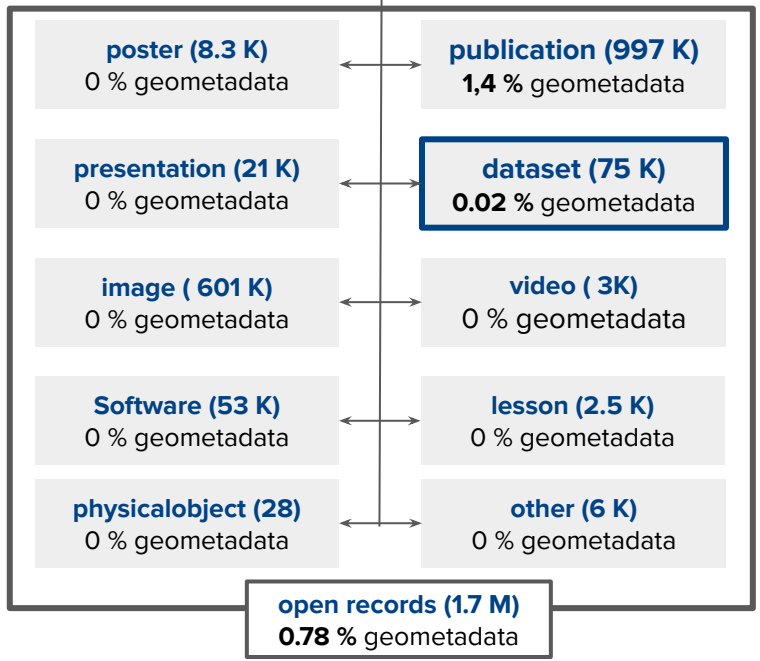
Case study:

zenodo

+1.7 M
Open records

10
types of records

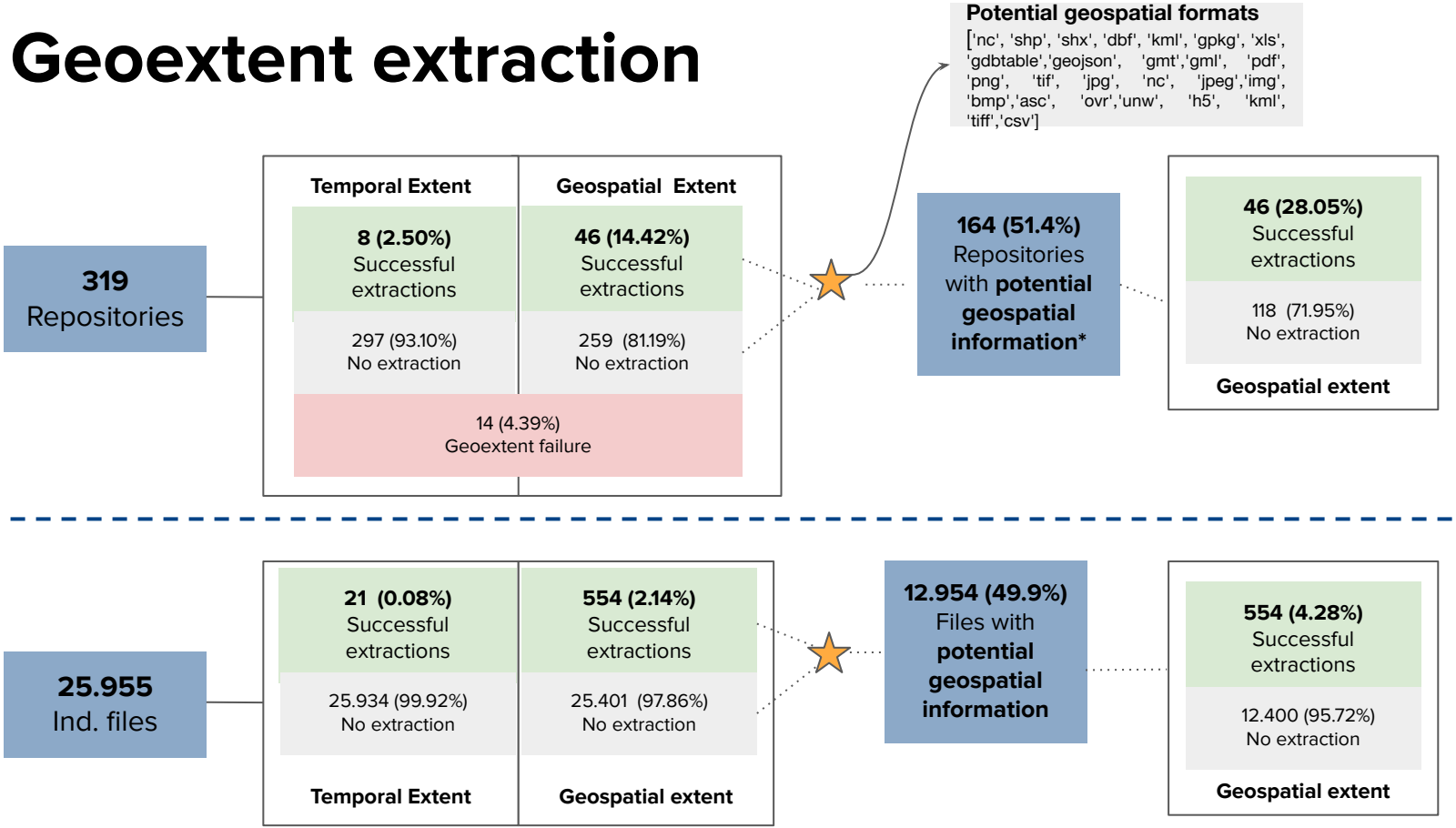
Zenodo metadata includes a geospatial ‘locations’ feature
(Not required)



Can we improve the metadata of zenodo ‘dataset’ records by using geoextent ?

Zenodo API
search term [geo&geology]
Max size: 500 MB
Sample size: 319 repositories

Geoextent extraction



potential geospatial information* = at least one file with geospatial format

Extraction by geospatial format

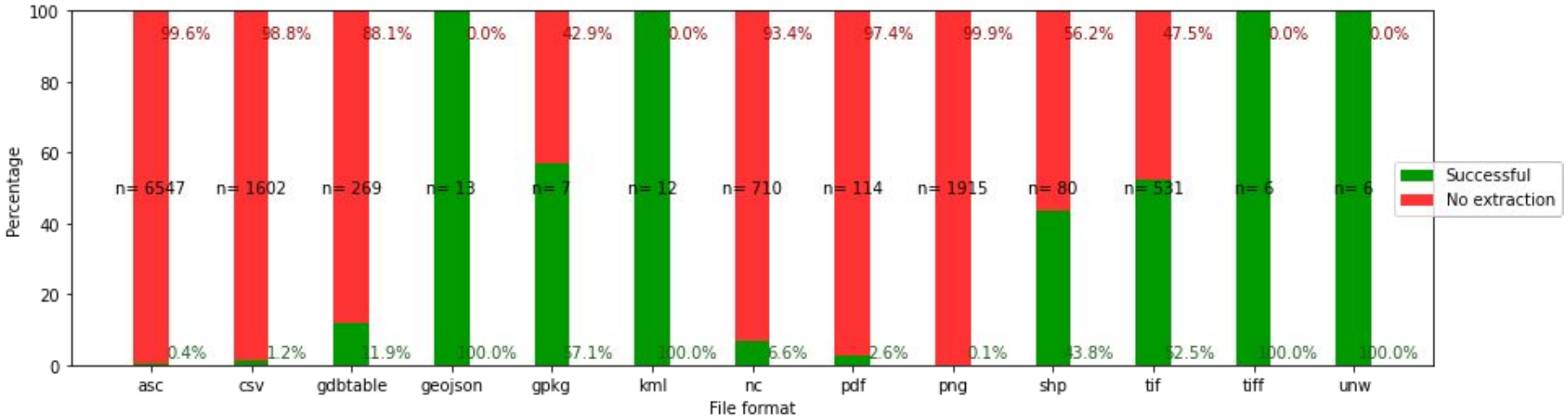


Figure 7. Geoextent extraction success by supported file type

https://o2r.info/geoextent/

Ambiguous formats	
asc	**csv**
png	pdf

Specialised formats				
nc	tif / tiff		kml*	
gbtable	**shp**	**gpkg***	geojson*	unw*

* small sample

Increasing success rate

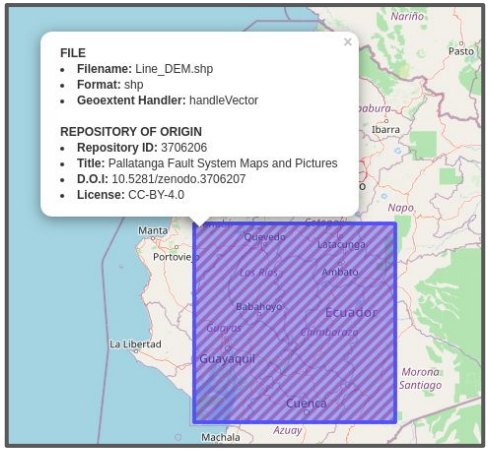
‘Successful’ extractions

dataset (75 K records)
0.02 % geometadata

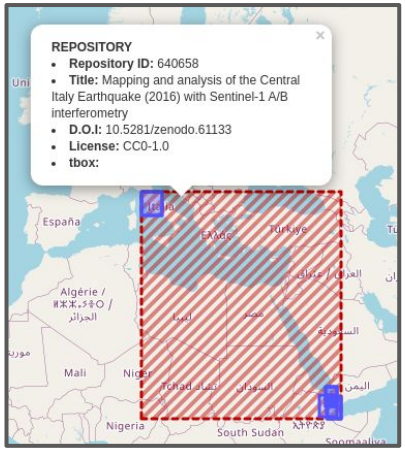
zenodo

Case study (319 records)
14.42% geometadata

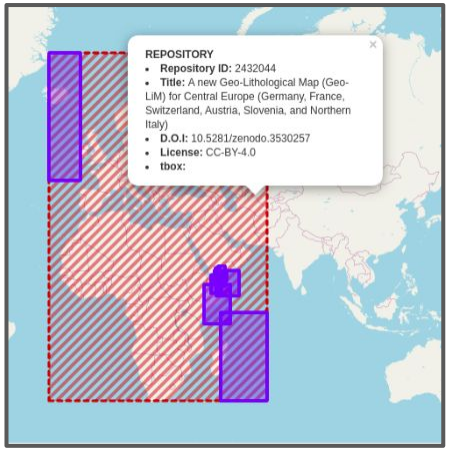
geoextent



Correct extraction*



Partially correct*



Incorrect*

*Based on human verification

Conclusions / ideas

Research records contain files with potential geospatial information

51.4% of zenodo records (n=319) contain geofiles formats

Ambiguous file formats are being used to store geospatial information.

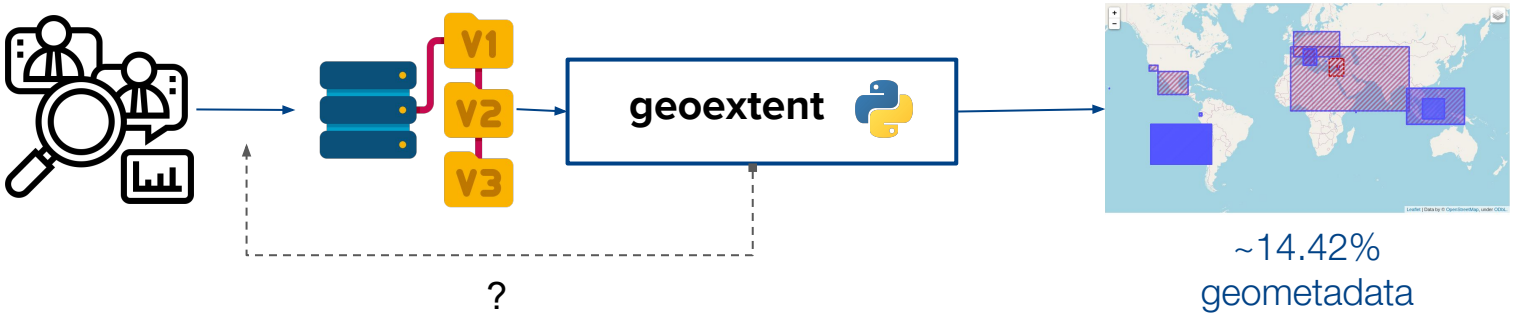
asc csv pdf png

Usage of specialized geospatial formats improve the preservation of information but do not completely guarantee it.

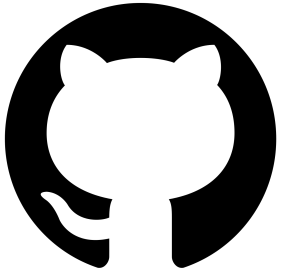
NetCDF: 6.6% ext. success rate
shapefile: 43.8 % ext. success rate

geoextent could be used to improve the rate of geometadata of existing records. Additional human verification could be needed.

Zenodo metadata: 0.2%
Geoextent: 14.4%



More about geoextent



geoextent development
https://github.com/o2r-project/geoextent/

New Issue

launch binder

More about reproducible research

o2r

Opening Reproducible Research
https://o2r.info/

DFG PE 1632/17-1