Using Free and Open Source GIS to Automatically Create Standards-Based Spatial Metadata

Claire Ellul

University College London

- Overview
  - The Problem with Metadata
  - Automation
  - Results
  - Further Work

The Problem with Metadata

Metadata

- Is “data about data”
- Gives you important information such as
  - When the data was created
  - Who by
  - For what purpose
  - When it was updated
  - How to obtain the data

The Problem with Metadata

However ..

….. Metadata is boring!

The Problem with Metadata

and Metadata is:

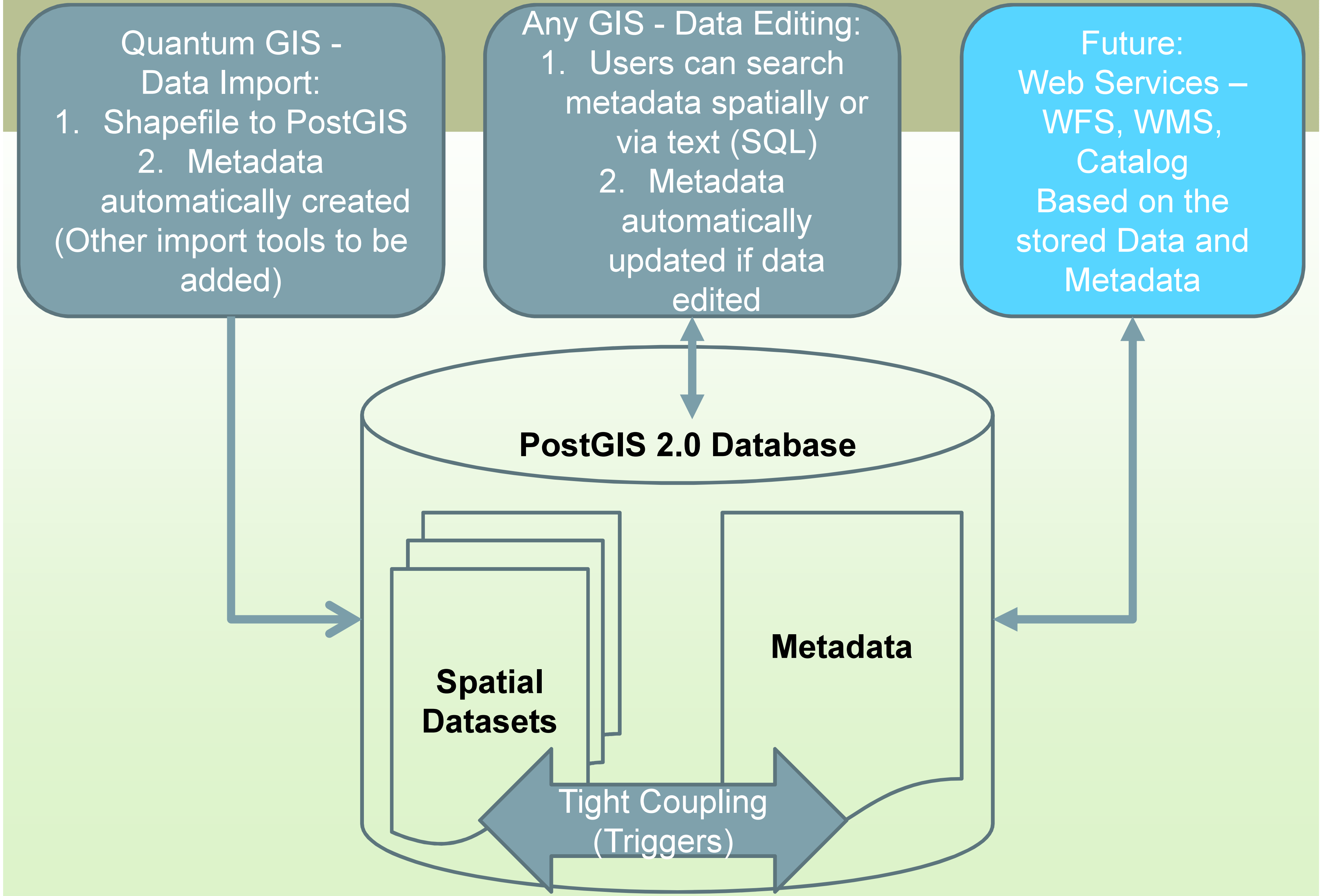
- Complex and time consuming to create
- Requires expertise about the data
- Requires expertise about how to create useful metadata
    - How much detail should be included?
    - Who are the end users of the metadata?
- Requires MAINTENANCE when data changes!

Metadata Automation

- Many elements of standards-based metadata may potentially be automatically created including:
  - Keywords
  - Dataset language
  - Metadata language
- FOSS tools provide a great environment for this!

Metadata Element	Automation Potential
Resource Title	Created manually. If not inserted by the user, default value is the dataset name (i.e. the PostGIS table name).
Resource Abstract	Created manually.
Resource Type	Can default to *dataset.* Automatically populated by PostGIS
Resource language	Can be automated using language detection algorithms
Keyword(s)	This could be implemented by concatenating all text fields of the dataset and picking the top 10 repeating words while eliminating common words.
Bounding Box	Can be automatically identified from the spatial coordinates in the dataset
Metadata language	This can be detected by applying a language detection algorithm to the metadata
Last Revision Date	Automatically update the metadata when the data changes

Metadata Element	Automation Potential
Metadata Date	Automatically defaults to the date the metadata was created/updated
Responsible Party	Can be populated automatically depending on the login (user id) for PostGIS
Metadata Contact	Can be populated automatically depending on the login (user id) for PostGIS
Resource Identifier	Can be automatically generated using the metadata record ID PostgreSQL identifiers
Metadata Geometry	Automatically created as a spatial geometry in PostGIS



Quantum GIS -
Data Import:

1. Shapefile to PostGIS
2. Metadata automatically created

(Other import tools to be added)

Any GIS - Data Editing:

1. Users can search metadata spatially or via text (SQL)
2. Metadata automatically updated if data edited

Future:
Web Services – WFS, WMS, Catalog
Based on the stored Data and Metadata

PostGIS 2.0 Database

Spatial Datasets

Metadata

Tight Coupling (Triggers)

Triggers in PostGIS

```
CREATE OR REPLACE FUNCTION public.add_boundingbox()
  RETURNS trigger AS $boundingbox$
-- this trigger function calculates the bounding box (Xmin, Xmax, Ymin, Ymax) of a
new dataset added to the database and inserts it in the metadata table
Declare
        table_name text; --variable that holds the name of the table (i.e. dataset)
        the_coord real; -- used to store the long/lat values
        curs1 refcursor; -- used to hold SQL query results
Begin
        …. the trigger code goes here …
End;

$boundingbox$ LANGUAGE plpgsql VOLATILE; -- VOLATILE indicates that the
                                            function value can change
```

Triggers in PostGIS

```
Open curs1 FOR EXECUTE
        'SELECT ST_XMax(ST_Extent(ST_Transform(the_geom,4326))) as the_coord
        FROM  '|| table_name;
        FETCH curs1 into the_coord;
        EXECUTE 'UPDATE metadata
                        SET bb_eastbound_long = ' || the_coord ||'
                         WHERE dataset_name =  '|| quote_literal(table_name);
CLOSE curs1;
```

Triggers in PostGIS

A series of INSERT triggers are run every time a new metadata record is created:

```
CREATE TRIGGER add_boundingbox
  AFTER INSERT
  ON public.metadata
  FOR EACH ROW
  EXECUTE PROCEDURE public.add_boundingbox();
```

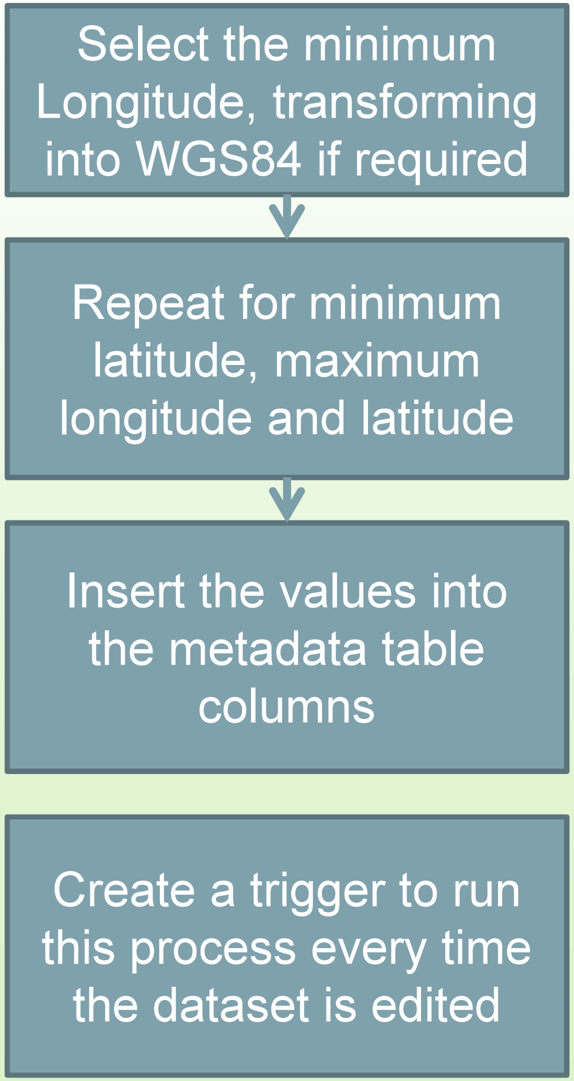
A series of metadata update triggers are run every time a dataset is modified, e.g.:

```
CREATE TRIGGER roaddata_bb_update
AFTER INSERT OR UPDATE OR DELETE on roaddata
FOR EACH ROW EXECUTE PROCEDURE update_bounding_box_roaddata();
```

*** NB: A new version of this trigger function is created automatically when a new spatial dataset is inserted into the database.*

Add Bounding Box

(PL/pgSQL Trigger)



Select the minimum Longitude, transforming into WGS84 if required

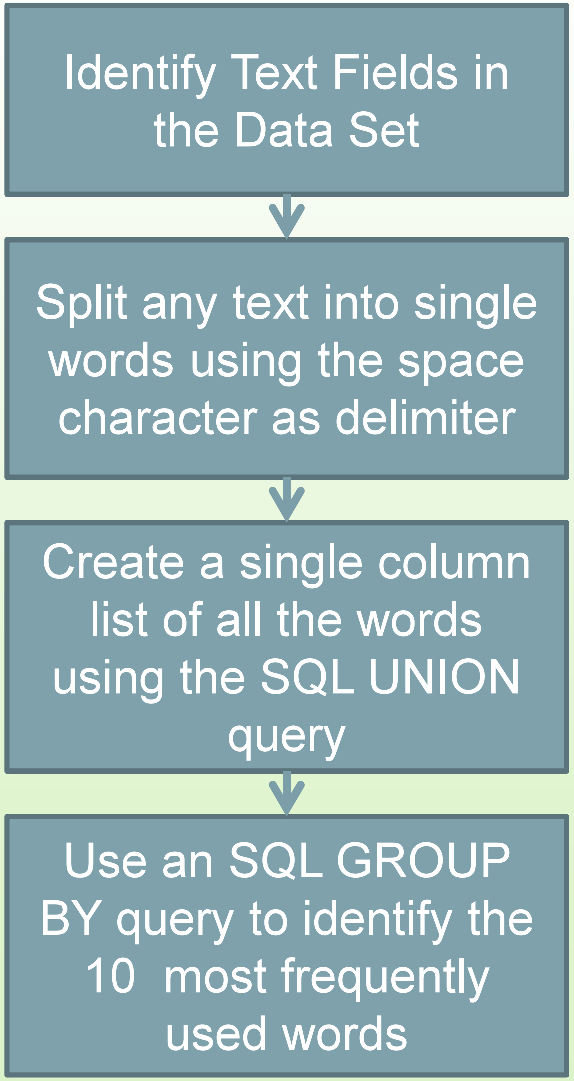
Repeat for minimum latitude, maximum longitude and latitude

Insert the values into the metadata table columns

Create a trigger to run this process every time the dataset is edited

Identifying Keywords

(PL/pgSQL Trigger)



Identify Text Fields in the Data Set

Split any text into single words using the space character as delimiter

Create a single column list of all the words using the SQL UNION query

Use an SQL GROUP BY query to identify the 10 most frequently used words

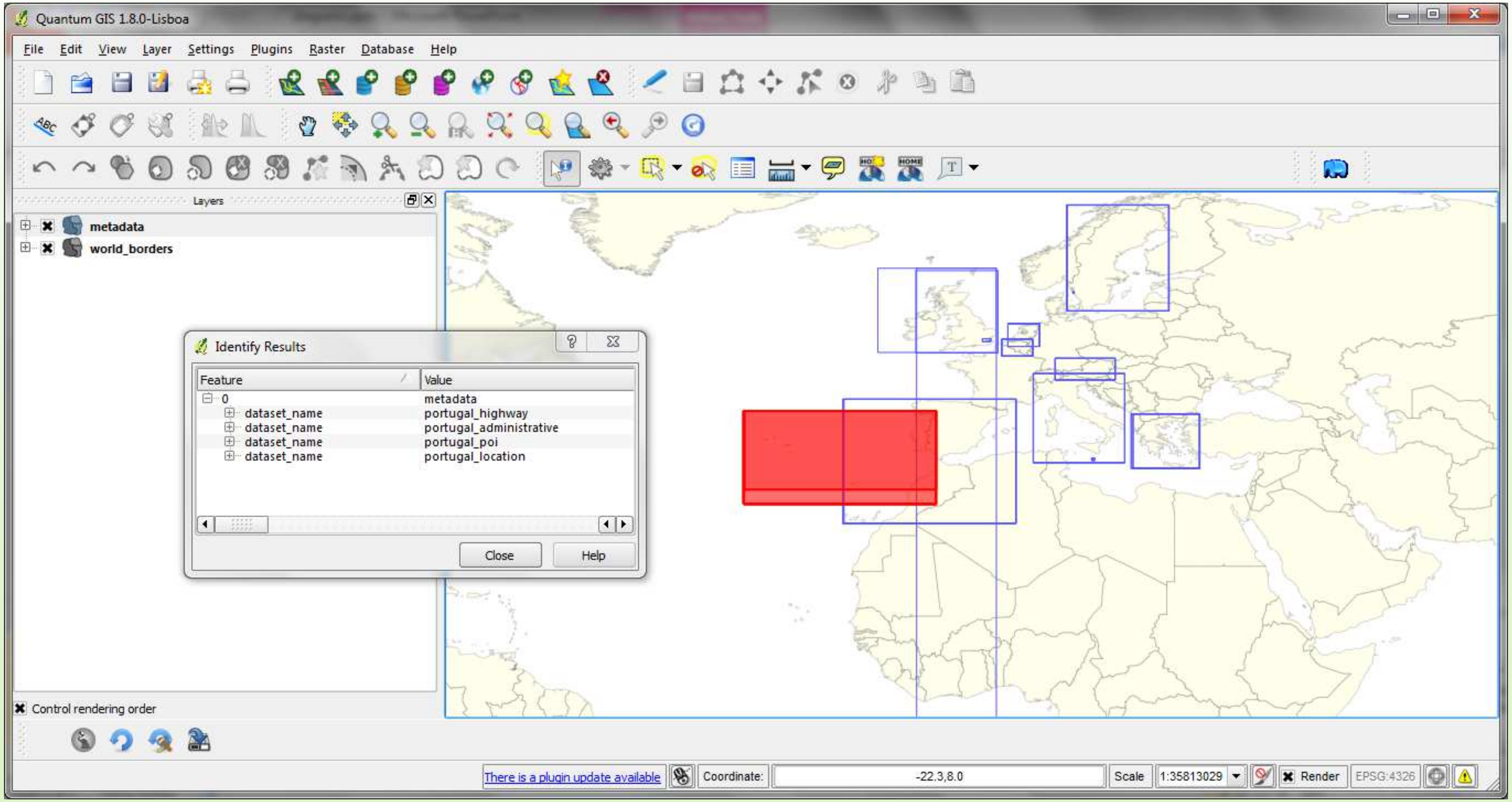
Testing the System

Metadata was created for Open Street Map datasets:

- Points of interest, administrative boundaries, road network and location data

And for 10 European countries

- UK, Austria, Greece, Malta, Italy, Spain, Belgium, Netherlands, Portugal, Sweden



Quantum GIS 1.8.0-Lisboa

Layers:
- ☒ metadata
- ☒ world_borders

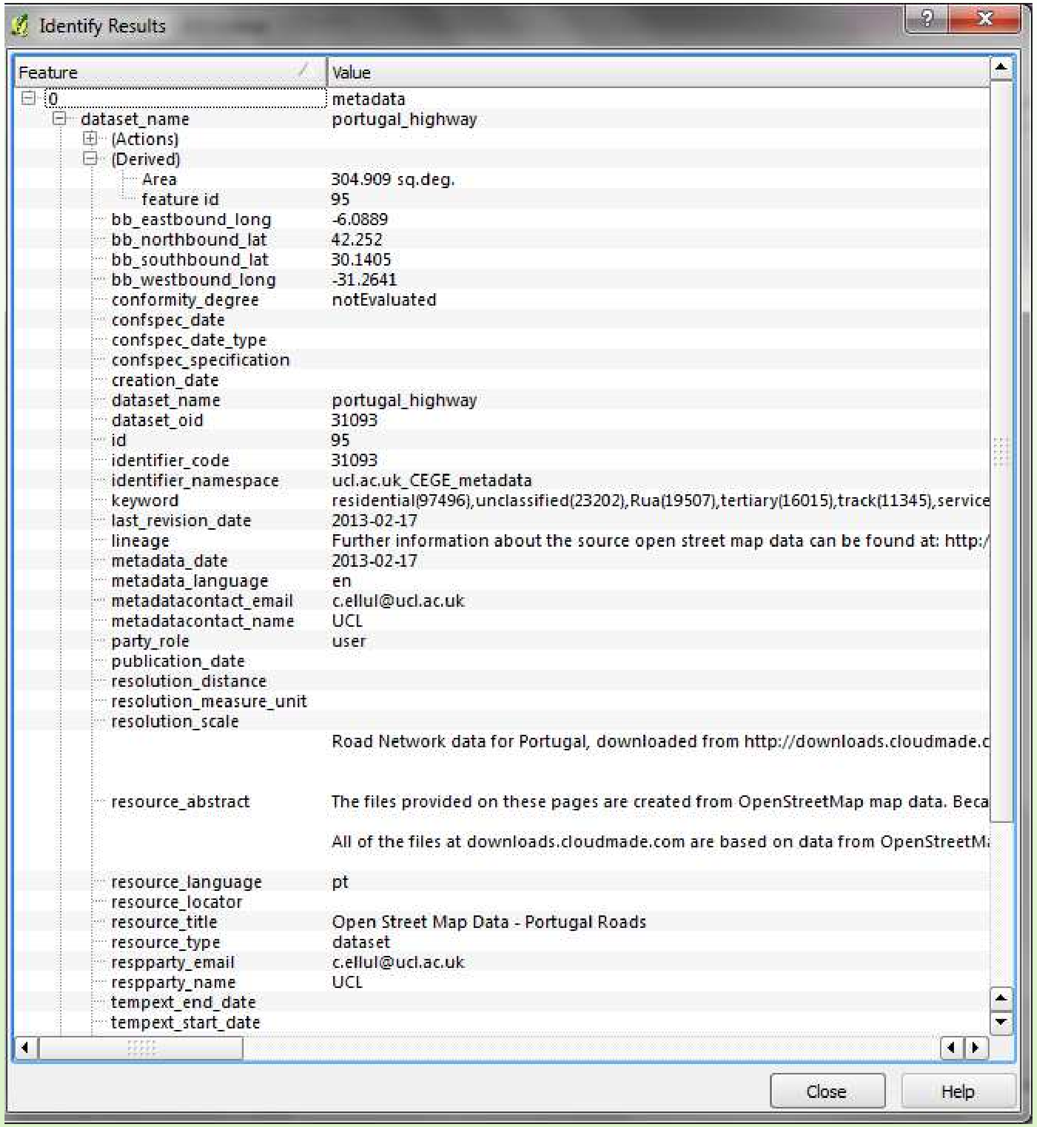
Identify Results

Feature	Value
0	metadata
dataset_name	portugal_highway
dataset_name	portugal_administrative
dataset_name	portugal_poi
dataset_name	portugal_location

Coordinate: -22.3,8.0

Scale 1:35813029

EPSG:4326



Feature	Value
0	metadata
dataset_name	portugal_highway
(Actions)	
(Derived)	
Area	304.909 sq.deg.
feature id	95
bb_eastbound_long	-6.0889
bb_northbound_lat	42.252
bb_southbound_lat	30.1405
bb_westbound_long	-31.2641
conformity_degree	notEvaluated
confspec_date	
confspec_date_type	
confspec_specification	
creation_date	
dataset_name	portugal_highway
dataset_oid	31093
id	95
identifier_code	31093
identifier_namespace	ucl.ac.uk_CEGE_metadata
keyword	residential(97496),unclassified(23202),Rua(19507),tertiary(16015),track(11345),service
last_revision_date	2013-02-17
lineage	Further information about the source open street map data can be found at: http:/
metadata_date	2013-02-17
metadata_language	en
metadatacontact_email	c.ellul@ucl.ac.uk
metadatacontact_name	UCL
party_role	user
publication_date	
resolution_distance	
resolution_measure_unit	
resolution_scale	
resource_abstract	Road Network data for Portugal, downloaded from http://downloads.cloudmade.c The files provided on these pages are created from OpenStreetMap map data. Beca All of the files at downloads.cloudmade.com are based on data from OpenStreetM:
resource_language	pt
resource_locator	
resource_title	Open Street Map Data - Portugal Roads
resource_type	dataset
respparty_email	c.ellul@ucl.ac.uk
respparty_name	UCL
tempext_end_date	
tempext_start_date	

Points of Interest – Keywords

- Keywords were in English and included:
    - ‘Public’, ‘Services’, ‘Tourism’, ‘Tree’, ‘Automotive’

Location Data – Keywords

- Keywords were predominantly in English and included ‘locality’ , ‘hamlet’, ‘village’
    - Also included place names: Aachen, Birmingham, Munchen, Trento

Administrative Areas – Keywords

- For keywords, the datasets yielded numbers such as 8, 6, 9, 10 in some cases

Road Network – Keywords

- Keywords were predominantly in English and included ‘track’ , ‘footway’, ‘cycleway’
  - Also included the words for ‘street’ in other languages: ‘via’ (Italian) ‘calle’ (Spanish), ‘strasse’ (Austria)

Summary Results

- Using FOSS (in particular PostGIS) means that the resulting data and metadata can be accessed from other GIS packages
- Metadata is automatically updated when data is modified
    - No matter which software is used to edit the data
- Using a central database means that the data and metadata can be published via OGC services such as WFS and Catalog Service for Web

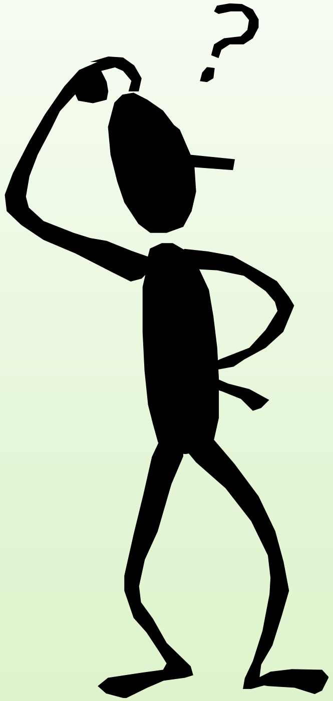
Further Work

- Testing with additional, more appropriate, single language datasets from different sources
- Extending the system to allow metadata to be created automatically for ANY spatial data in a PostGIS database, no matter how it is loaded
- Publishing the data and metadata via tools such as GeoServer

Further Work

- Identify and resolve any issues related to performance – i.e. the time taken to create the metadata each time the data is modified.
- Improve handling of non-Latin character-sets
- Thinking about deployment – how to ensure that the approach can be used by users not having spatial database expertise

Any Questions?



c.ellul@ucl.ac.uk

Metadata Element	Automation Potential
Resource Title	Created manually. If not inserted by the user, default value is the dataset name.
Resource Abstract	Created manually.
Resource Type	Can default to dataset
Resource language	Can be automated using language detection algorithms
Keyword(s)	This could be implemented by concatenating all text fields of the dataset and picking the top 10 repeating words while eliminating common words.
Bounding Box	Can be automatically identified from the spatial coordinates in the dataset
Date of publication	Can default to the date that data was uploaded to the system, with updates when the data is edited. Manual verification required by the end user.
Date of last revision	Default to the date the data was uploaded to the system. Update automatically any time data edited

Metadata Element	Automation Potential
Date of creation	Default to the date the data was uploaded. Manual verification required by the end user
Limitations on public access and conditions of use (2 elements)	Given the academic context, a default value can be assigned, perhaps taking the most open value or perhaps on a per project basis.
Responsible party name, email and role (3 elements)	Based on user groups (identified from the user’s login details and a corresponding lookup table).
Metadata contact name, email and date (3 elements)	This can be derived from the database login of the person uploading the dataset or creating the new dataset.
Metadata language	This can be detected by applying a language detection algorithm to the metadata

The Problem with Metadata

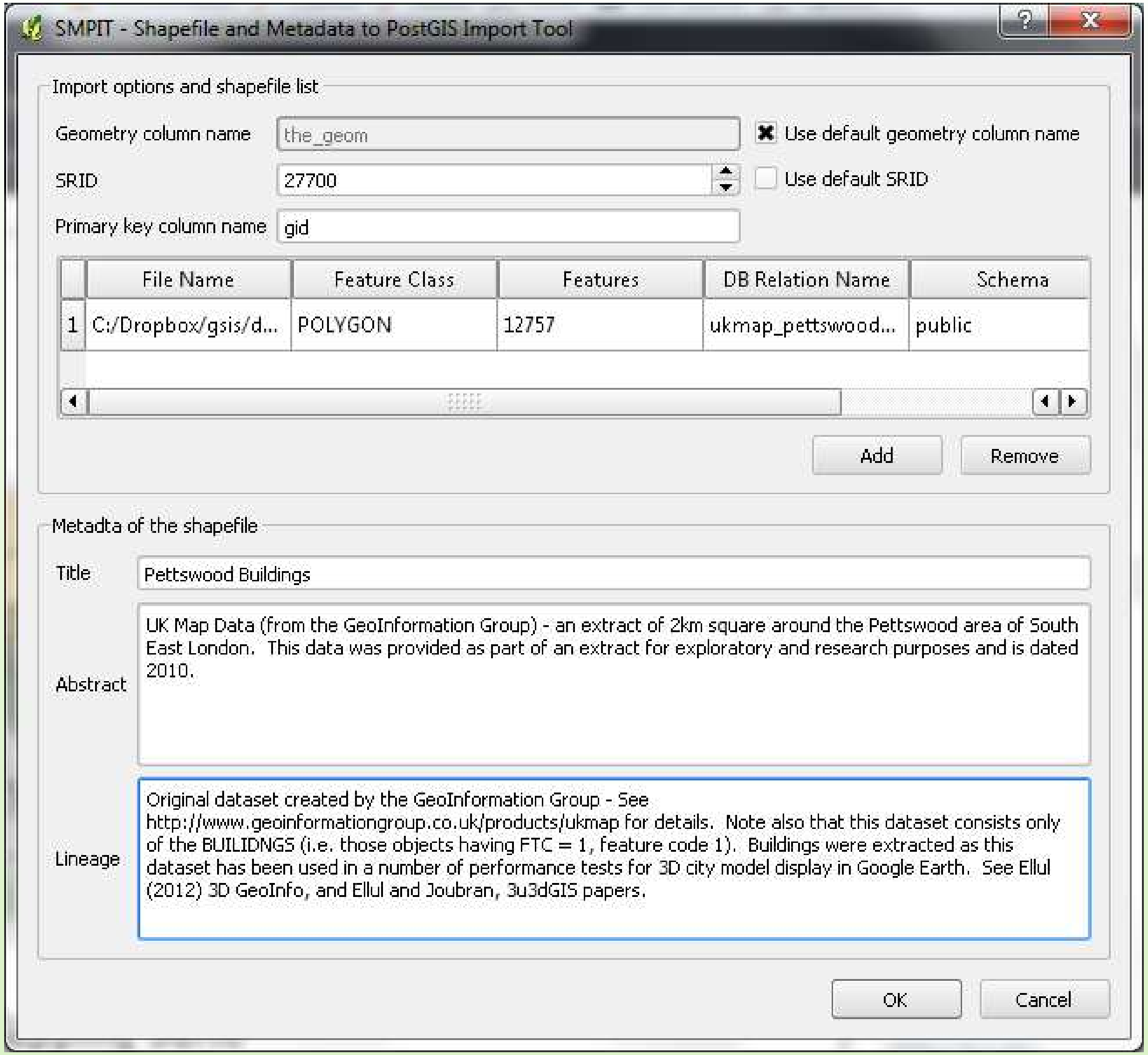
Metadata is important for academic research:

- The EU FP7 project SECOA is developing models of coastal conflicts in countries including Italy, India, the UK, Portugal, Israel, Vietnam, Sweden and Belgium
- Data underpins these models

The Problem with Metadata

Metadata is important for academic research:

- However, comparable data is not always available so alternative data is sometimes substituted
    - In the Italian case, the total “Employees in industrial sector in Rome” is not available. “**The number of employees was not available so the number of local units in the industrial sector has been used”**
- Without this metadata, the SECOA team could be comparing employees with industrial units!



Import options and shapefile list

Geometry column name: the_geom

☒ Use default geometry column name

SRID: 27700

□ Use default SRID

Primary key column name: gid

	File Name	Feature Class	Features	DB Relation Name	Schema
1	C:/Dropbox/gsis/d...	POLYGON	12757	ukmap_pettswood...	public

Metadta of the shapefile

Title: Pettswood Buildings

Abstract: UK Map Data (from the GeoInformation Group) - an extract of 2km square around the Pettswood area of South East London. This data was provided as part of an extract for exploratory and research purposes and is dated 2010.

Lineage: Original dataset created by the GeoInformation Group - See http://www.geoinformationgroup.co.uk/products/ukmap for details. Note also that this dataset consists only of the BUILIDNGS (i.e. those objects having FTC = 1, feature code 1). Buildings were extracted as this dataset has been used in a number of performance tests for 3D city model display in Google Earth. See Ellul (2012) 3D GeoInfo, and Ellul and Joubran, 3u3dGIS papers.