

Automating Metadata Creation

Closer than we think

Jo Cook, Astun Technology

Hello! I'm Jo *waves*. I'm the Technical Evangelist for Data Discovery at Astun Technology.



<http://astuntechnology.com> | @astuntech

Speaker notes

The job title is just a fancy way of saying that I help people find and share data. By the way, this image was taken pre-pandemic when it was OK to stand so close together, and indeed to get haircuts. Astun provide web-based mapping (and related) tools and training in the UK, based on the open source geospatial stack.



Speaker notes

I like to find and fix little blockers, for example by making documentation easier to follow or teaching people to use GitHub

I'm not a coder! My passion is enabling others to do their job as easily as possible, preferably with open source tools.



Speaker notes

I like to find and fix little blockers, for example by making documentation easier to follow or teaching people to use GitHub

I'm not a coder! My passion is enabling others to do their job as easily as possible, preferably with open source tools.



Recently my focus has been on metadata, and I've learnt the following important fact...

Speaker notes

I like to find and fix little blockers, for example by making documentation easier to follow or teaching people to use GitHub

Metadata is hard!



Speaker notes

While people understand the need for metadata, they often don't get any training, and it's an adjunct to their day job, documentation is very technical and so on and so on...

Metadata is hard!

- It's complex and time-consuming

Speaker notes

While people understand the need for metadata, they often don't get any training, and it's an adjunct to their day job, documentation is very technical and so on and so on...

Metadata is hard!

- It's complex and time-consuming
- It's hard to know where to start

Speaker notes

While people understand the need for metadata, they often don't get any training, and it's an adjunct to their day job, documentation is very technical and so on and so on...

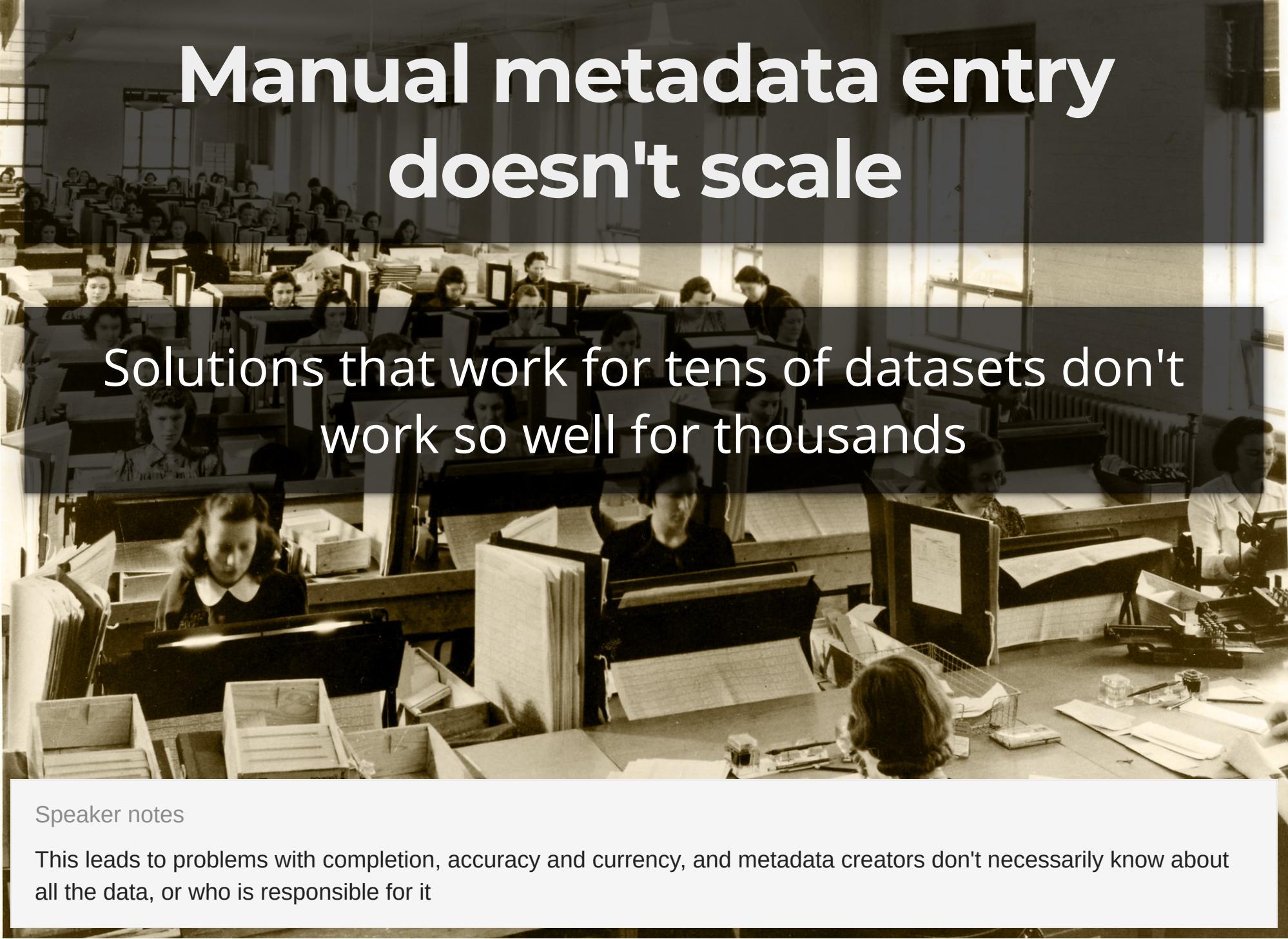
Metadata is hard!

- It's complex and time-consuming
- It's hard to know where to start
- There's a steep learning curve

Speaker notes

While people understand the need for metadata, they often don't get any training, and it's an adjunct to their day job, documentation is very technical and so on and so on...

Manual metadata entry doesn't scale

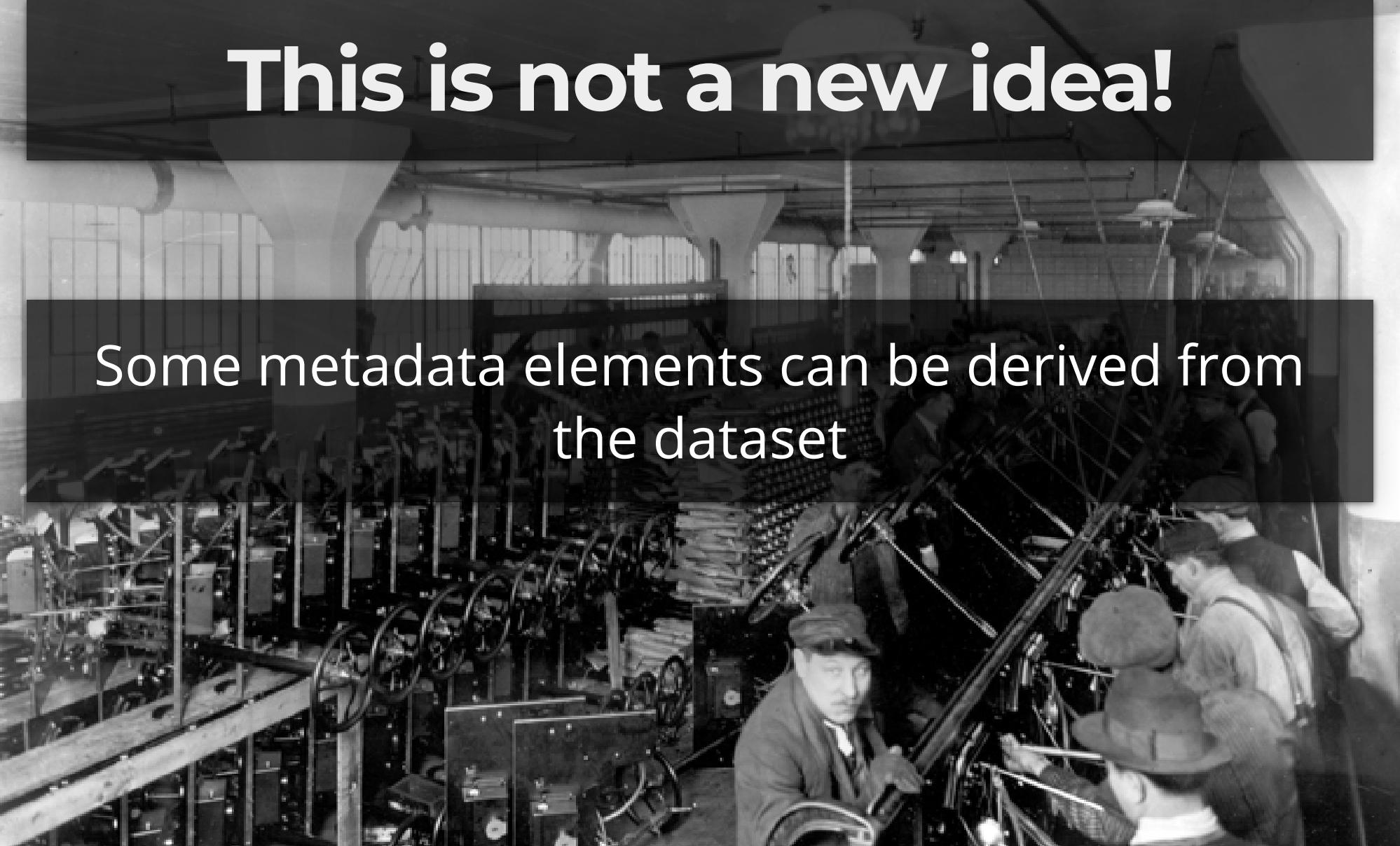


Solutions that work for tens of datasets don't work so well for thousands

Speaker notes

This leads to problems with completion, accuracy and currency, and metadata creators don't necessarily know about all the data, or who is responsible for it

This is not a new idea!



Some metadata elements can be derived from the dataset

Speaker notes

There are many tools out there that will derive at least some metadata elements for you, including the USGS Metadata Wizard, which has been around for many years

Using FOSS is not even very original!

Free and Open Source Geospatial (FOSS4G)
Conference Proceedings

Volume 13 Nottingham, UK

Article 8

2013

There are many open source geospatial libraries
that can be used to derive some elements
Academia - First Investigations

Claire Ellul

University College London, United Kingdom

Nart Tamash

University College London, United Kingdom

Feng Xian

University College London, United Kingdom

Speaker notes

GDAL/OGR, JTS, GEOS, Boost..., not only in geospatial but for things like CityML



Layer Properties - high_water_polyline | Information

Information from provider

Name	high_water_polyline
Path	/home/jo/Documents/geodata/borders/high_water_polyline.shp
Storage	ESRI Shapefile
Comment	
Encoding	ISO-8859-1
Geometry	Line (MultiLineString)
CRS	EPSG:27700 - OSGB 1936 / British National Grid Projected
Extent	5512.998499999996944,5336.9660000000003492 : 655653.849999999767169,122334.008500991942
Unit	meters
Feature count	47,357

Identification

Identifier	
Parent Identifier	
Title	
Type	dataset
Language	
Abstract	
Categories	
Keywords	

Extent

CRS

Spatial Extent

Most approaches cover the elements that are easy to programmatically derive, such as extents, titles, creation or update dates, language and so on

Speaker notes

From personal observation, it's these manual elements that cause the most trouble and cause metadata projects to flounder or sometimes fail all together. These are increasingly important for data discovery and sharing (eg linked data)



□

Layer Properties - high_water_polyline | Information

Information from provider

Name	high_water_polyline
Path	/home/jo/Documents/geodata/boundaryline/Data/GB/high_water_polyline.shp
Storage	ESRI Shapefile
Comment	
Encoding	ISO-8859-1
Geometry	Line (MultiLineString)
CRS	EPSG:27700 - OSGB 1936 / British National Grid - Projected
Extent	5512.998499999996944,5336.9660000000003492 : 655653.849999999767169,1220301.5020000000949949
Unit	meters
Feature count	47,357

Identification

Identifier	
Parent Identifier	
Title	
Type	dataset
Language	
Abstract	
Categories	
Keywords	

Extent

CRS	Spatial Extent
-----	----------------

Speaker notes

From personal observation, it's these manual elements that cause the most trouble and cause metadata projects to flounder or sometimes fail all together. These are increasingly important for data discovery and sharing (eg linked data)



Layer Properties - high_water_polyline | Information

Information from provider

Name	high_water_polyline
Path	/home/jo/Documents/geodata/borders/high_water_polyline.shp
Storage	ESRI Shapefile
Comment	
Encoding	ISO-8859-1
Geometry	Line (MultiLineString)
CRS	EPSG:27700 - OSGB 1936 / British National Grid Projected
Extent	5512.998499999996944,5336.966000000003492 : 655653.849999999767169,122.33333333333333
Unit	meters
Feature count	47,357

Identification

Identifier	
Parent Identifier	
Title	
Type	dataset
Language	
Abstract	
Categories	
Keywords	

Extent

CRS	Spatial Extent
-----	----------------

But there are other elements that need a human touch, such as human-readable alternative titles, abstracts and keywords that can't be derived in the same way

Speaker notes

From personal observation, it's these manual elements that cause the most trouble and cause metadata projects to flounder or sometimes fail all together. These are increasingly important for data discovery and sharing (eg linked data)

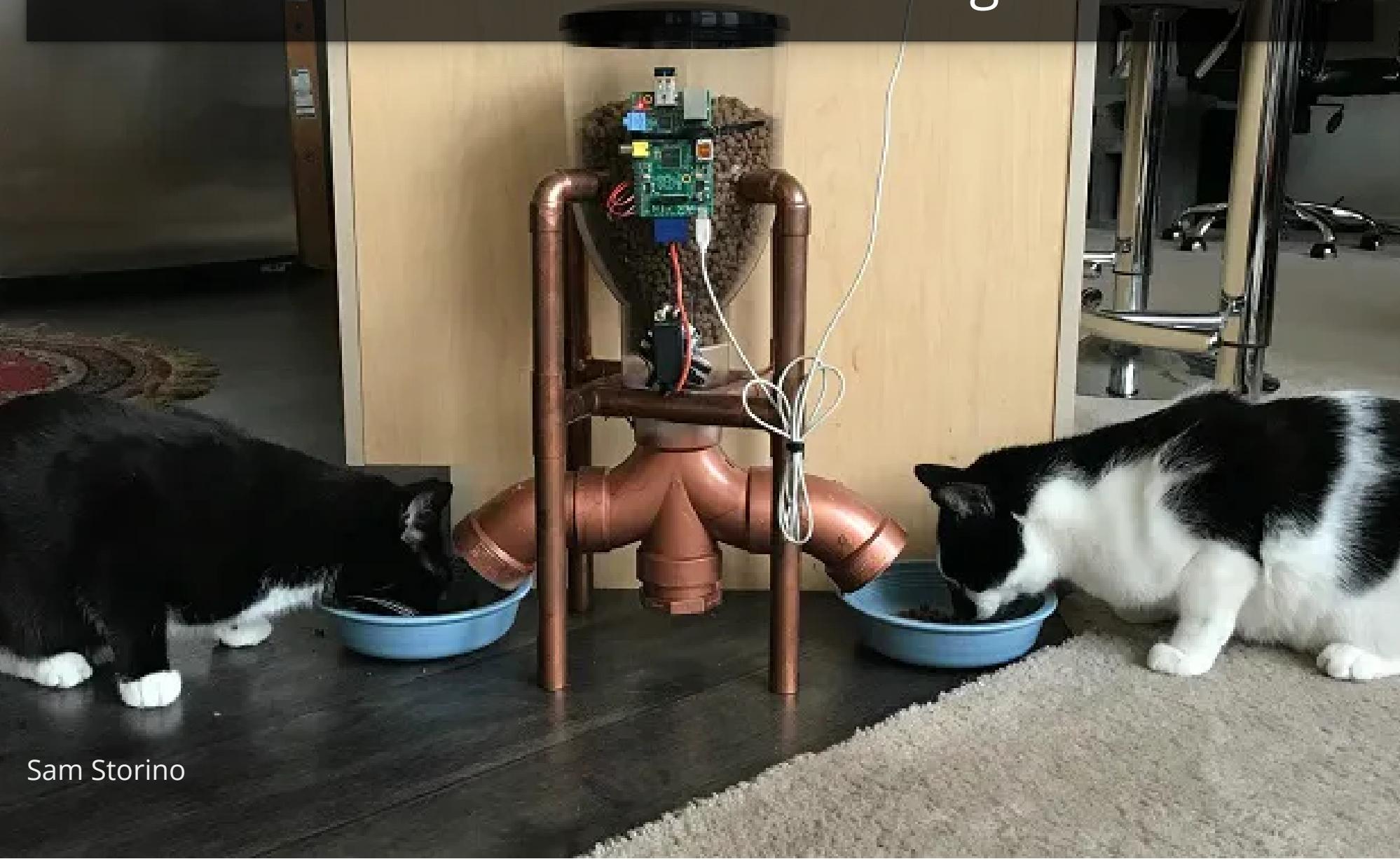
We're developing a solution that links together a number of open source technologies to overcome the "manual challenge"



Speaker notes

Some of it is very proof of concept at the moment so I don't have a slick demo

Our solution comprises a number of open source tools and libraries linked together



Sam Storino

Part One: Metadata Crawler

Crawler is a script for discovering spatial data (vector or raster) in file systems and databases. For each data source it finds, it derives as much of the metadata as it can, or inserts place-holder elements for others.



Speaker notes

Crawler was created by Titellus, using Talend ETL's spatial plugin. It's open source. It can be run as a web service (using tomcat) or as a cross-platform shell script

STEP 1: FROM DATA TO PLACEHOLDER METADATA



```
<?xml version="1.0"  
encoding="UTF-8"?>  
<gmd:MD_Metadata>  
...  
</gmd:MD_Metadata>
```

e,placeholderorg;placeholdercity...

Speaker notes

Not just Geonetwork. It can update existing records. For UK metadata, we've modified Crawler so that it outputs metadata in Gemini 2.3 format, and we're looking at modifying it to create non-spatial metadata too

STEP 1: FROM DATA TO PLACEHOLDER METADATA



```
<?xml version="1.0"  
encoding="UTF-8"?>  
<gmd:MD_Metadata>  
...  
</gmd:MD_Metadata>
```

Crawler creates an ISO19139 metadata record for each data source it discovers, and an ISO19110 Feature Catalog record for vector data

e,placeholderorg;placeholdercity...

Speaker notes

Not just Geonetwork. It can update existing records. For UK metadata, we've modified Crawler so that it outputs metadata in Gemini 2.3 format, and we're looking at modifying it to create non-spatial metadata too

STEP 1: FROM DATA TO PLACEHOLDER METADATA



```
<?xml version="1.0"  
encoding="UTF-8"?>  
<gmd:MD_Metadata>  
...  
</gmd:MD_Metadata>
```

It can create static XML, or insert records directly into any Metadata Catalog with a CSW-T (Transactional CSW) endpoint.

e,placeholderorg;placeholdercity...

Speaker notes

Not just Geonetwork. It can update existing records. For UK metadata, we've modified Crawler so that it outputs metadata in Gemini 2.3 format, and we're looking at modifying it to create non-spatial metadata too

Table geonetwork.notpublic.docker_notpublicpolygon

abstract needs to be 100 characters long or more, to satisfy Gemini 2.3 so here are a few more characters so that it's long enough

On going

Download and links



PlaceHolderURL

PlaceHolderDesc <https://www.centralbedfordshire.gov.uk/>

[Open link](#)



notpublic.docker_notpublicpolygon

This dataset is published in the database

PG:dbname=geonetwork host=172.17.0.2 port=5432 user=****
password=**** in table notpublic.docker_notpublicpolygon.

Feature catalog



Feature catalog geonetwork.notpublic.docker_notpublicpolygon

Technical information

Name	Definition
------	------------

name (String)	
---------------	--

[Feature catalog](#)

Overview



No ratings

See all feedback

Add your review

Spatial extent



Temporal extent

Creation date

2000-01-01

Revision date

2020-06-15

Period

Sat Jan 01 2000 00:00:00 GMT+0000 ➔ Sat Jan 01 2000 00:00:00 GMT+0000

Provided by

Table geonetwork.notpublic.docker_notpublicpolygon

abstract needs to be 100 characters long or more, to satisfy Gemini 2.3 so here are a few more characters so that it's long enough

On going

Download and links



PlaceHolderURL

PlaceHolderDesc <https://www.centralbedfordshire.gov.uk/>



notpublic.docker_notpublicpolygon

This dataset is published in the database

PG:dbname=geonetwork host=172.17.0.2 port=5432 user=****
password=**** in table notpublic.docker_notpublicpolygon.

Feature catalog



Feature catalog geonetwork.notpublic.docker_notpublicpolygon

Technical information

Name	Definition
------	------------

name (String)

Feature catalog

Overview

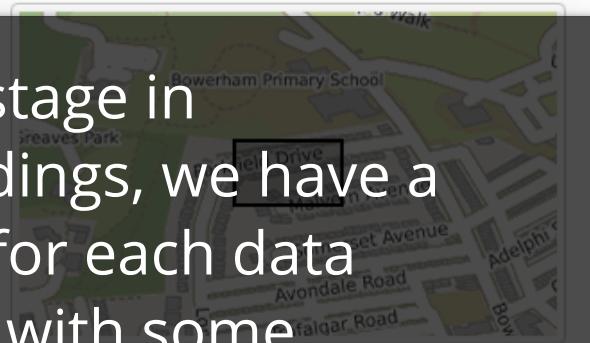


No ratings

See all feedback

Add your review

Spatial extent



Open link

At this stage in proceedings, we have a record for each data source, with some programmatic elements...

Creation date

2000-01-01

Revision date

2020-06-15

Period

Sat Jan 01 2000 00:00:00 GMT+0000 ➡ Sat Jan 01 2000 00:00:00 GMT+0000

Provided by

Table geonetwork.notpublic.docker_notpublicpolygon

abstract needs to be 100 characters long or more, to satisfy Gemini 2.3 so here are a few more characters so that it's long enough

On going

Download and links



PlaceHolderURL

PlaceHolderDesc <https://www.centralbedfordshire.gov.uk/>

Open link



notpublic.docker_notpublicpolygon

This dataset is published in the database
PG:dbname=geonetwork host=172.17.0.2 port=5432 user=****
password=**** in table notpublic.docker_notpublicpolygon.

Feature catalog



Feature catalog geonetwork.notpublic.docker_notpublicpolygon

Technical information

Name	Definition
------	------------

name (String)	
---------------	--

Feature catalog

Overview



No ratings

See all feedback

Add your review

Spatial extent

Temporal extent

...some placeholder elements, and a linked feature catalog record describing the attributes.

Creation date

2000-01-01

Revision date

2020-06-15

Period

Sat Jan 01 2000 00:00:00 GMT+0000 ➡ Sat Jan 01 2000 00:00:00 GMT+0000

Provided by

Table geonetwork.notpublic.docker_notpublicpolygon

abstract needs to be 100 characters long or more, to satisfy Gemini 2.3 so here are a few more characters so that it's long enough

On going

Download and links



PlaceHolderURL

PlaceHolderDesc <https://www.centralbedfordshire.gov.uk/>

Open link



notpublic.docker_notpublicpolygon

This dataset is published in the database

PG:dbname=geonetwork host=172.17.0.2 port=5432 user=****
password=**** in table notpublic.docker_notpublicpolygon.

Feature catalog



Feature catalog geonetwork.notpublic.docker_notpublicpolygon

Technical information

Name	Definition
------	------------

name (String)	
---------------	--

Feature catalog

Overview



No ratings

See all feedback

Add your review

Spatial extent



Temporal extent

Creation date

2000-01-01

Revision date

2020-06-15

Period

Sat Jan 01 2000 00:00:00 GMT+0000 ➡ Sat Jan 01 2000 00:00:00 GMT+0000

Provided by

Part Two: Ye Olde Approache



A	B	C	D	E
	abstract	keywords	updatefrequency	inspirekeyword
	<p>Descriptive text about the table.</p> <p>This should be at least 100 characters long to satisfy the Gemini 2.3 guidance. The purpose of an abstract is to provide a clear narrative summary that enables the reader or user to understand the content of the data. See the README sheet for information on the other fields to be completed</p>			
chema.table	completed	location	weekly	buildings
			continual	
			daily	
			weekly	
			fortnightly	
			monthly	

Speaker notes

Metadata Crawler derives titles using a set of rules (for databases it uses the form Database.Schema.Tablename) so we can use the same when auto-generating the rows in the excel spreadsheet to ensure a match. To ensure we get quality results from this process, we can use controlled text fields for elements where that is appropriate. This may seem clumsy but excel is well-known and used. Data owners can copy and paste in bulk to populate "their" records quickly and easily.

A	B	C	D	E
	abstract	keywords	updatefrequency	inspirekeyword
	<p>Descriptive text about the table.</p> <p>This should be at least 100 characters long to satisfy the Gemini 2.3 guidance. The purpose of an abstract is to provide a clear narrative summary that enables the reader or user to understand the content of the data. See the README sheet for information on the other fields to be completed</p>		We do a second run through the data to create a spreadsheet with a row per record for data owners to complete the manual fields, then use the GeoNetwork API wrapped in a python script to match spreadsheet rows to metadata records in the catalogues.	
chema.table		location	continual	
		daily	fortnightly	
		monthly	annually	
		catalogue	dataset	

Speaker notes

Metadata Crawler derives titles using a set of rules (for databases it uses the form Database.Schema.Tablename) so we can use the same when auto-generating the rows in the excel spreadsheet to ensure a match. To ensure we get quality results from this process, we can use controlled text fields for elements where that is appropriate. This may seem clumsy but excel is well-known and used. Data owners can copy and paste in bulk to populate "their" records quickly and easily.

STEP 2: METADATA FROM CSV

```
title,abstract,keywords,updatefrequency  
db.geo.buses,"this is a dataset about buses",buses,daily
```

```
host:8080/geonetwork/srv/eng/q?title  
db.geo.buses"
```

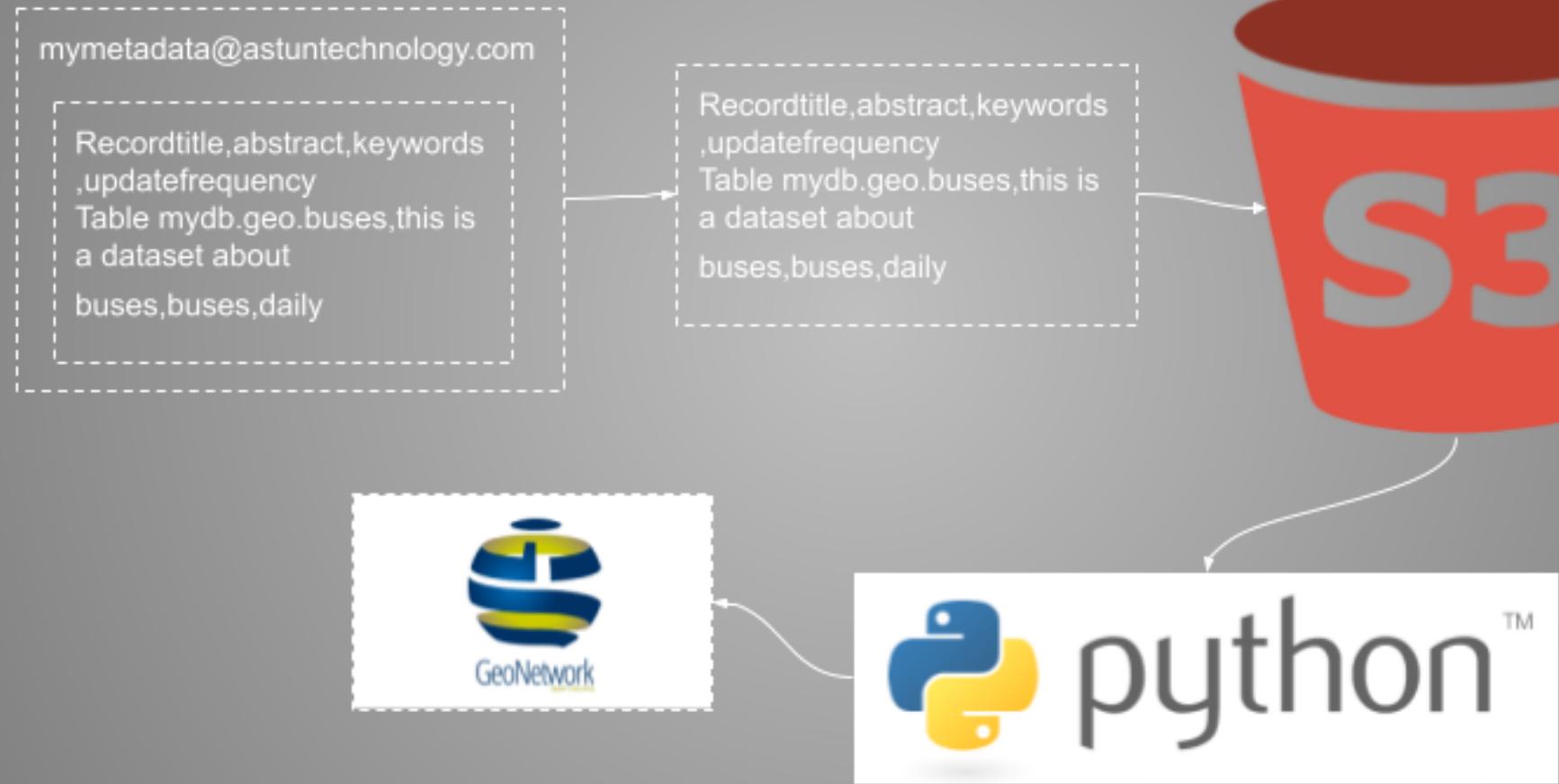


```
[  
  {  
    "uuid": "xxx"  
    "this is a dataset about buses" :  
    "<snip>/gmd:abstract/gco:CharacterString"  
  }  
]
```

Speaker notes

Even though GeoNetwork is hosted on AWS in the cloud, we can let users run these scripts using environments such as Cloud9

STEP THREE: EMAILING CSV

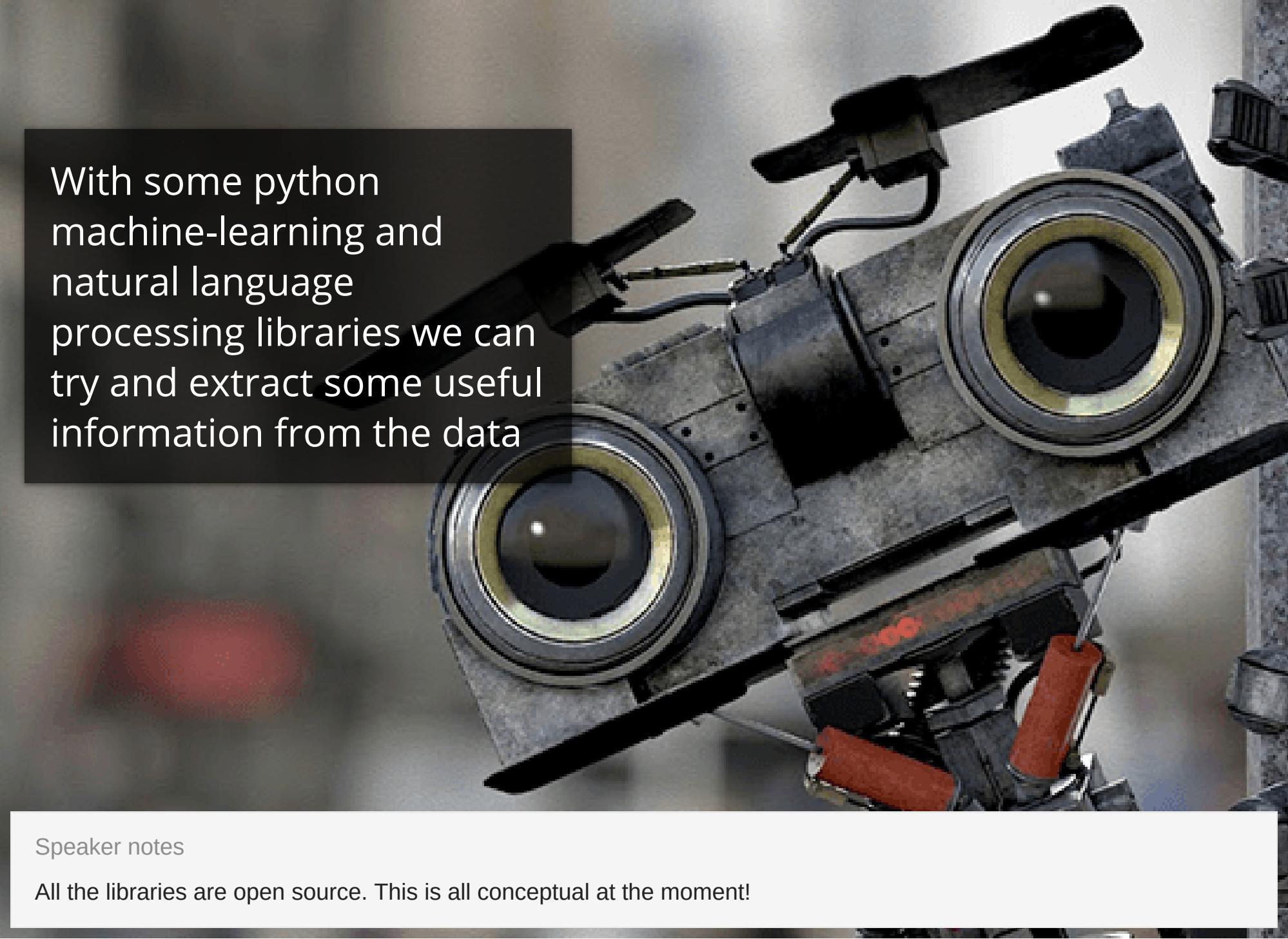


Speaker notes

For extra geek points, and to automate the workflow even more, we can allow people to send the spreadsheet as a csv to an email address associated with an S3 bucket

Part Two: The Cool New Experimental Approach

**BEYOND THIS POINT
YOU SHOULD ENGAGE
A GUIDE**

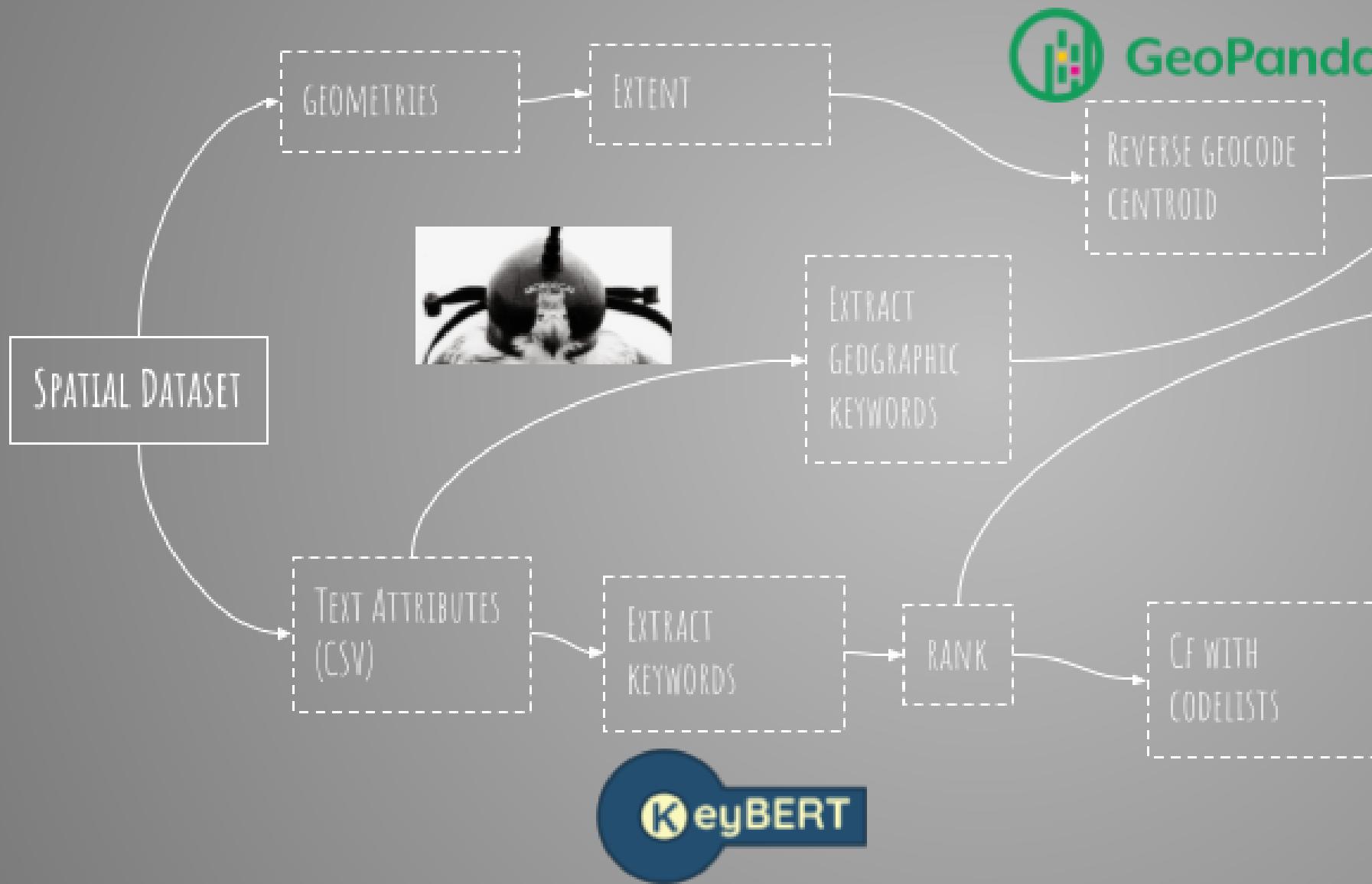


With some python
machine-learning and
natural language
processing libraries we can
try and extract some useful
information from the data

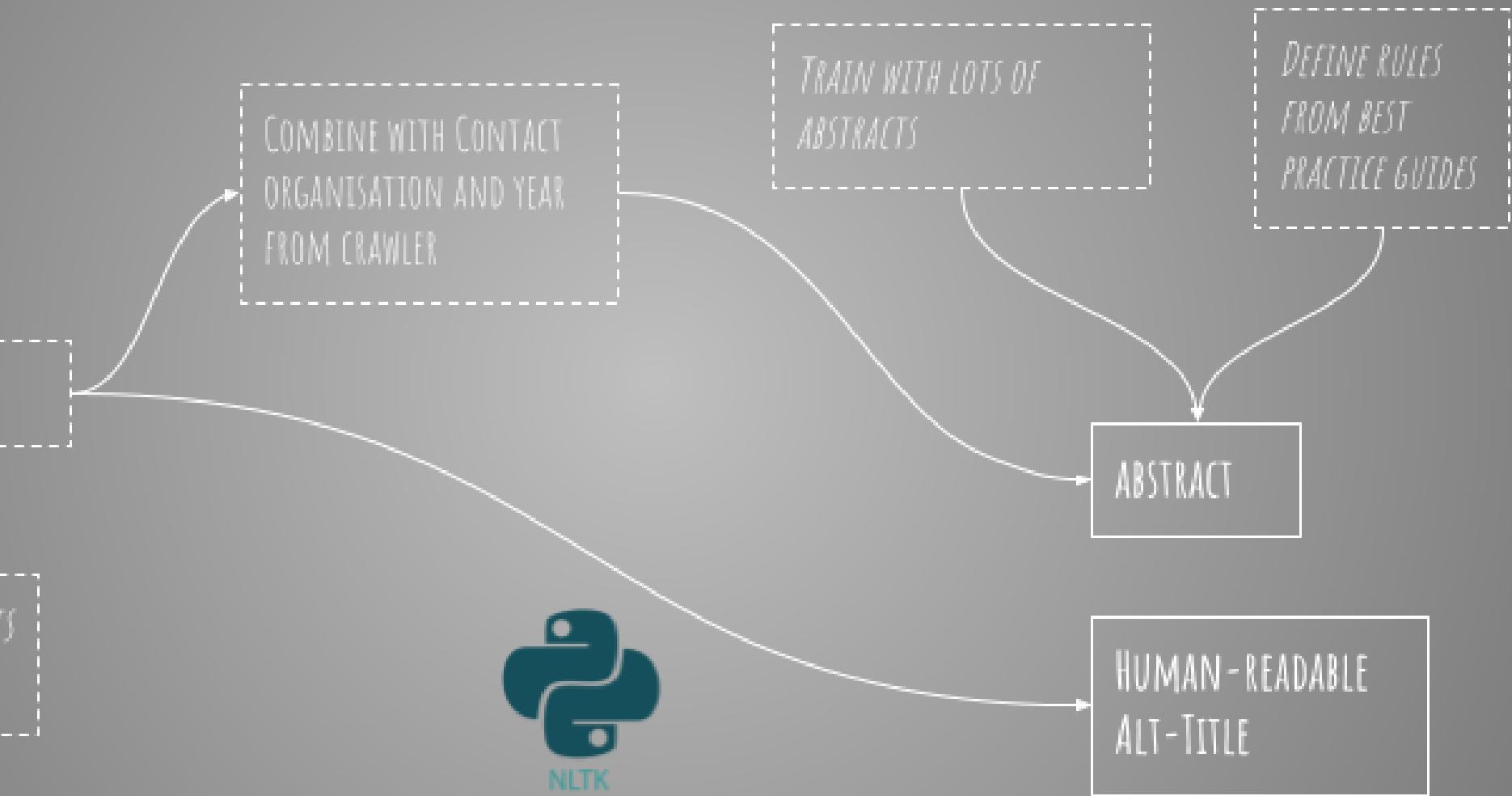
Speaker notes

All the libraries are open source. This is all conceptual at the moment!

MACHINE-LEARNING STEP ONE



MACHINE-LEARNING STEP TWO





Geospatial
Commission

Search engine optimisation (SEO) for data publishers: a quick guide

Speaker notes

The guidance is from an SEO perspective and indicates how the elements like titles and abstracts might be displayed in standard web search results. So (eg) the min and max length for an abstract



Geospatial
Commission

Search engine optimisation (SEO) for data publishers: a quick guide

We have a large corpus of metadata, and some best practice guidelines for data sharing to help train our model and define some rules.

Speaker notes

The guidance is from an SEO perspective and indicates how the elements like titles and abstracts might be displayed in standard web search results. So (eg) the min and max length for an abstract



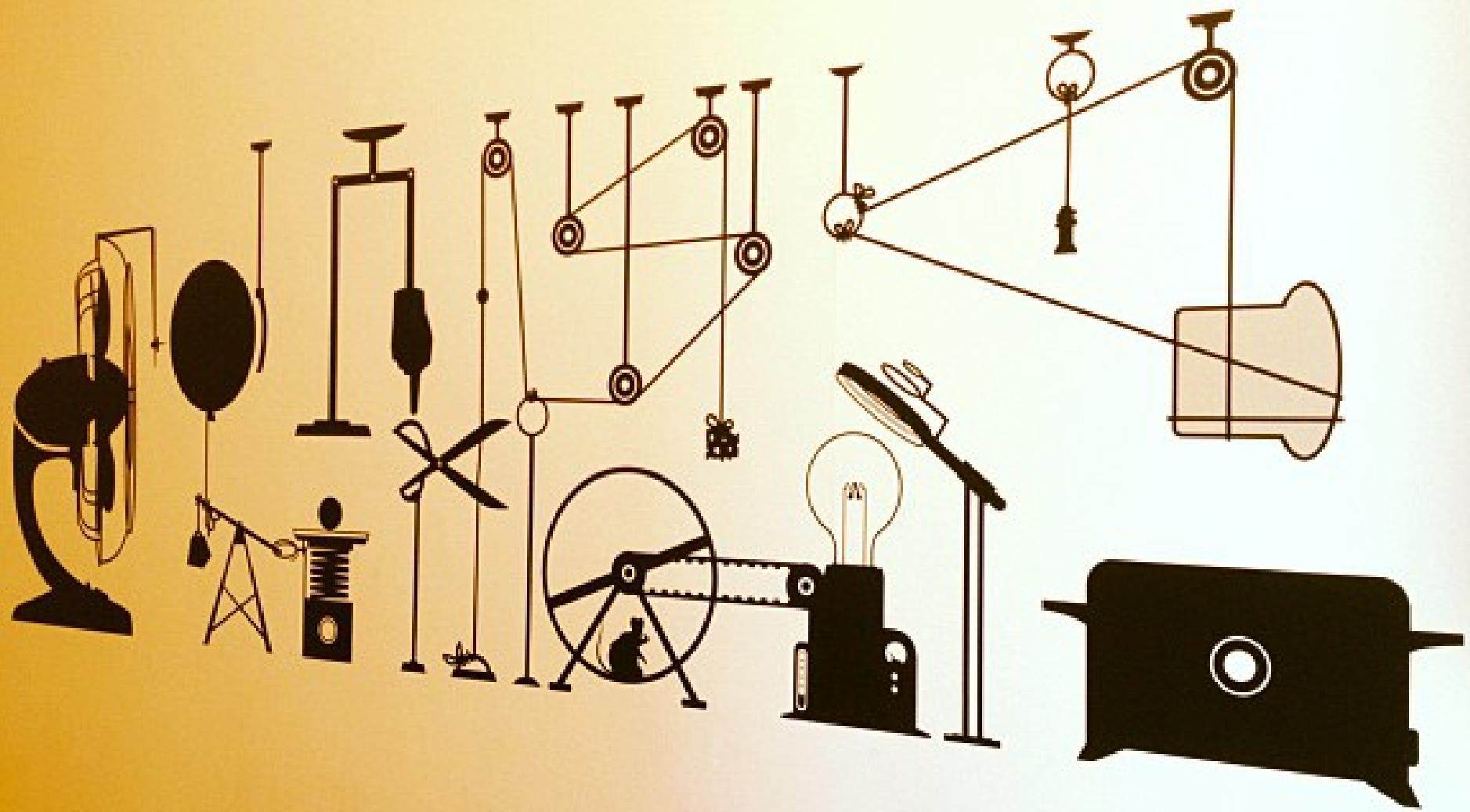
Geospatial
Commission

Search engine optimisation (SEO) for data publishers: a quick guide

Once we have these additional elements we can again use the GeoNetwork API in a python wrapper to update the metadata records

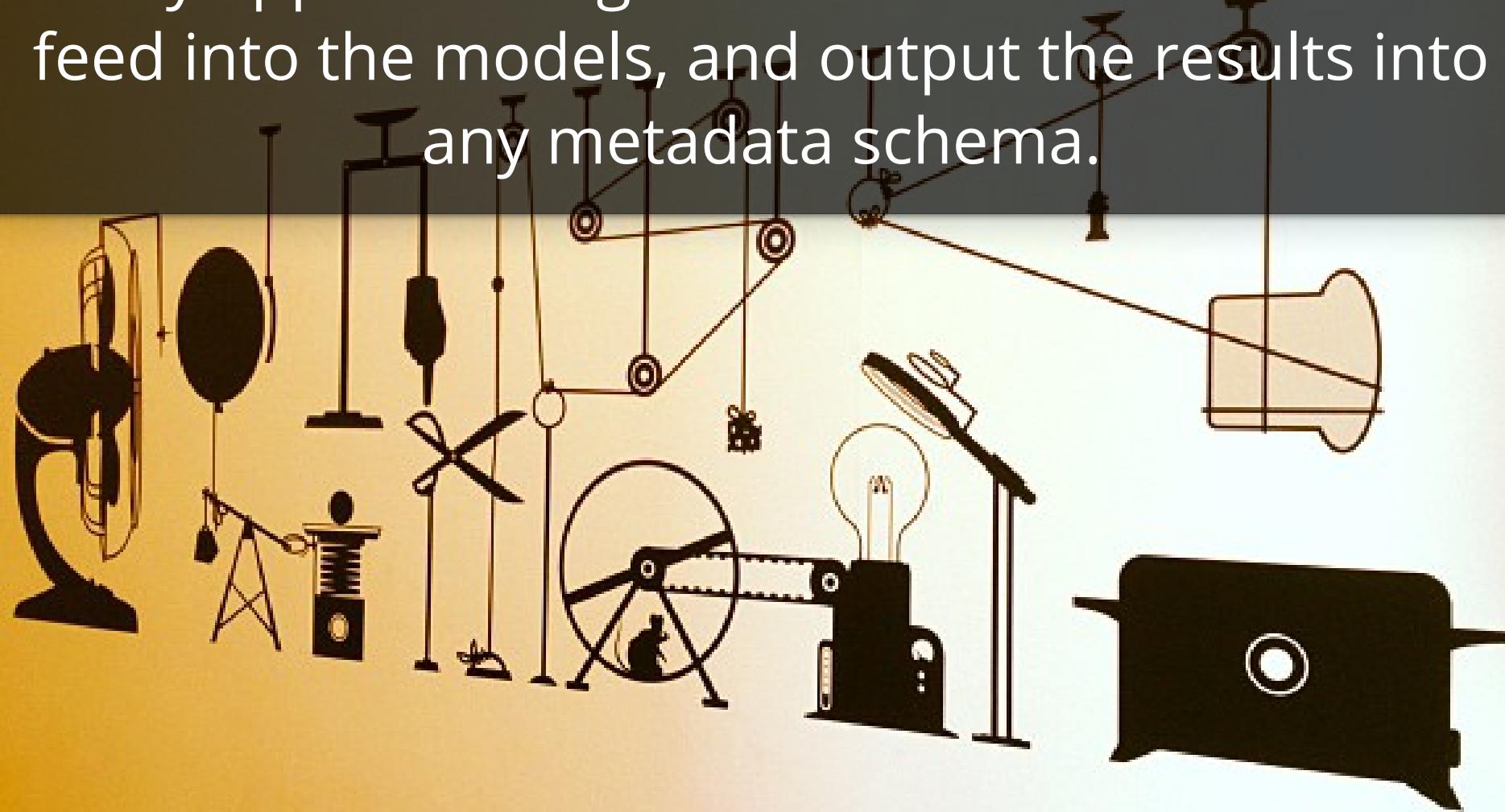
Speaker notes

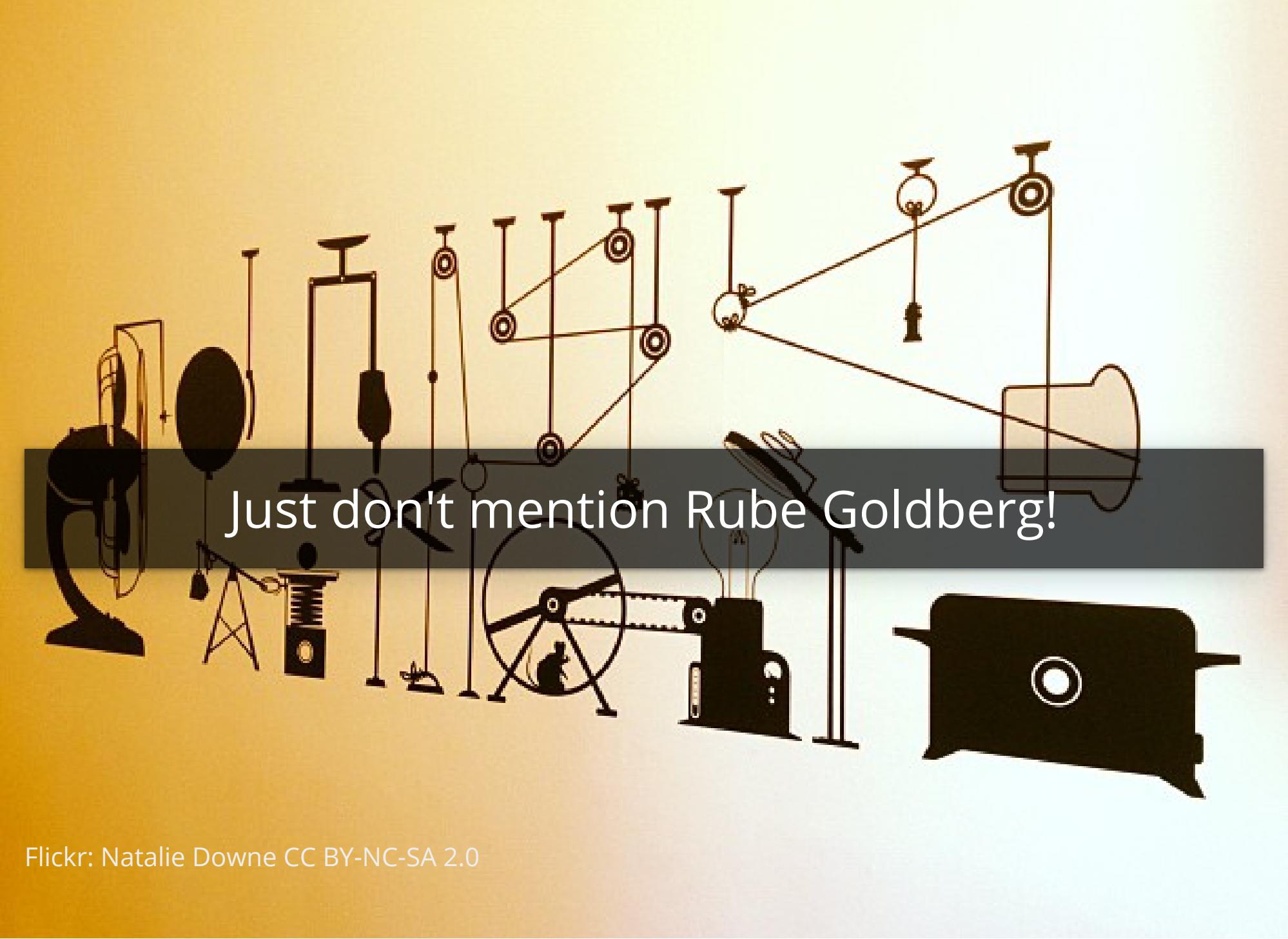
The guidance is from an SEO perspective and indicates how the elements like titles and abstracts might be displayed in standard web search results. So (eg) the min and max length for an abstract



Flickr: Natalie Downe CC BY-NC-SA 2.0

It's modular and schema-agnostic. You can use any approach to get the derived metadata to feed into the models, and output the results into any metadata schema.





Just don't mention Rube Goldberg!



The end result will be valid metadata records that need minimal human intervention

Speaker notes

These are designed to be a starting-point rather than fully completed records. Some human intervention is expected. In GeoNetwork at least, you can keep them up to date using harvesting (defining which elements should and should not be overwritten)

The usual caveats apply!



Useful References

Crawler | SEO for publishers | Natural Language
Toolkit | KeyBert | Mordecai | Geopandas

Thank You!

Jo Cook, Astun Technology

jocook@astuntechnology.com