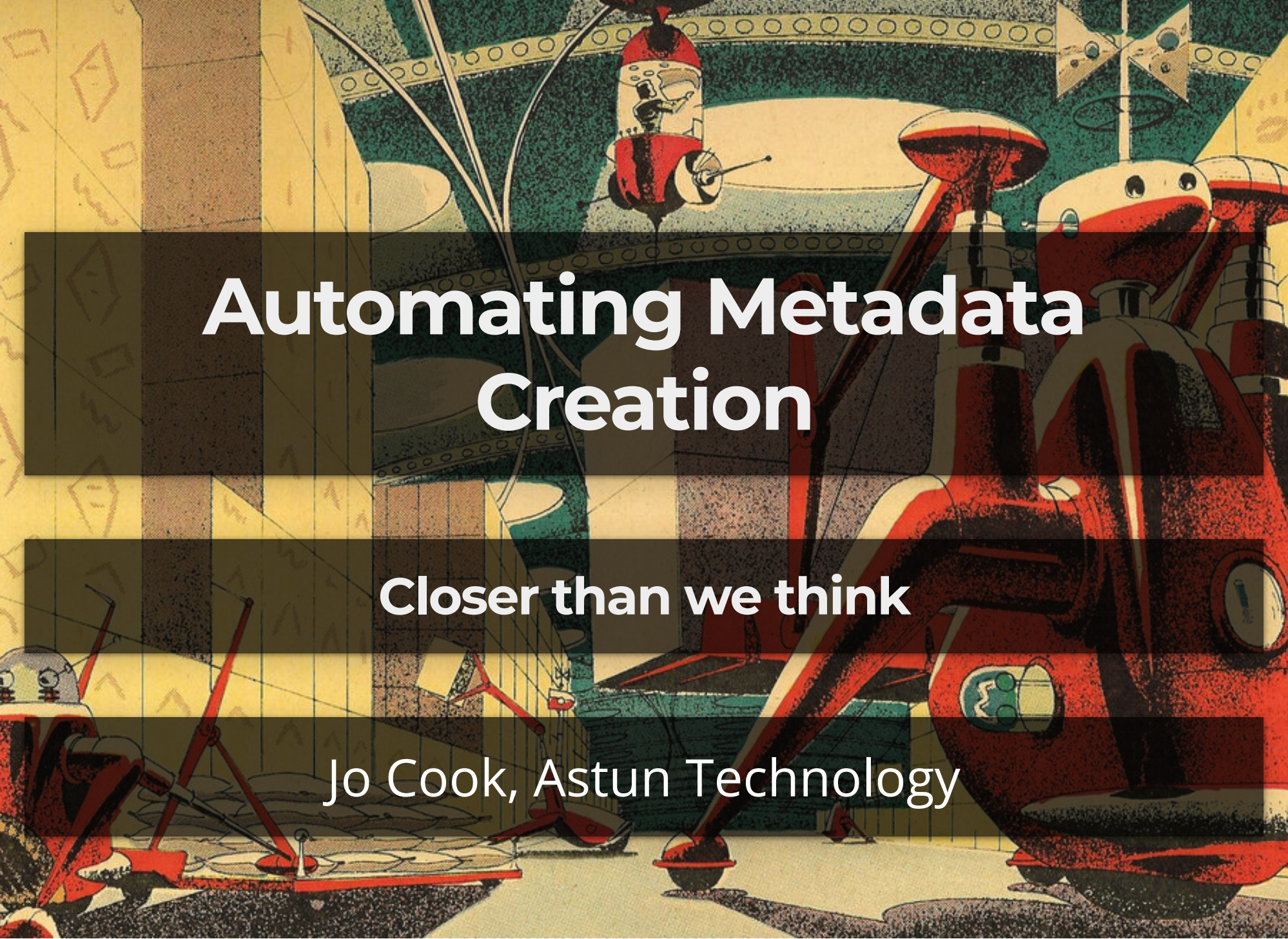


Automating Metadata Creation

Closer than we think

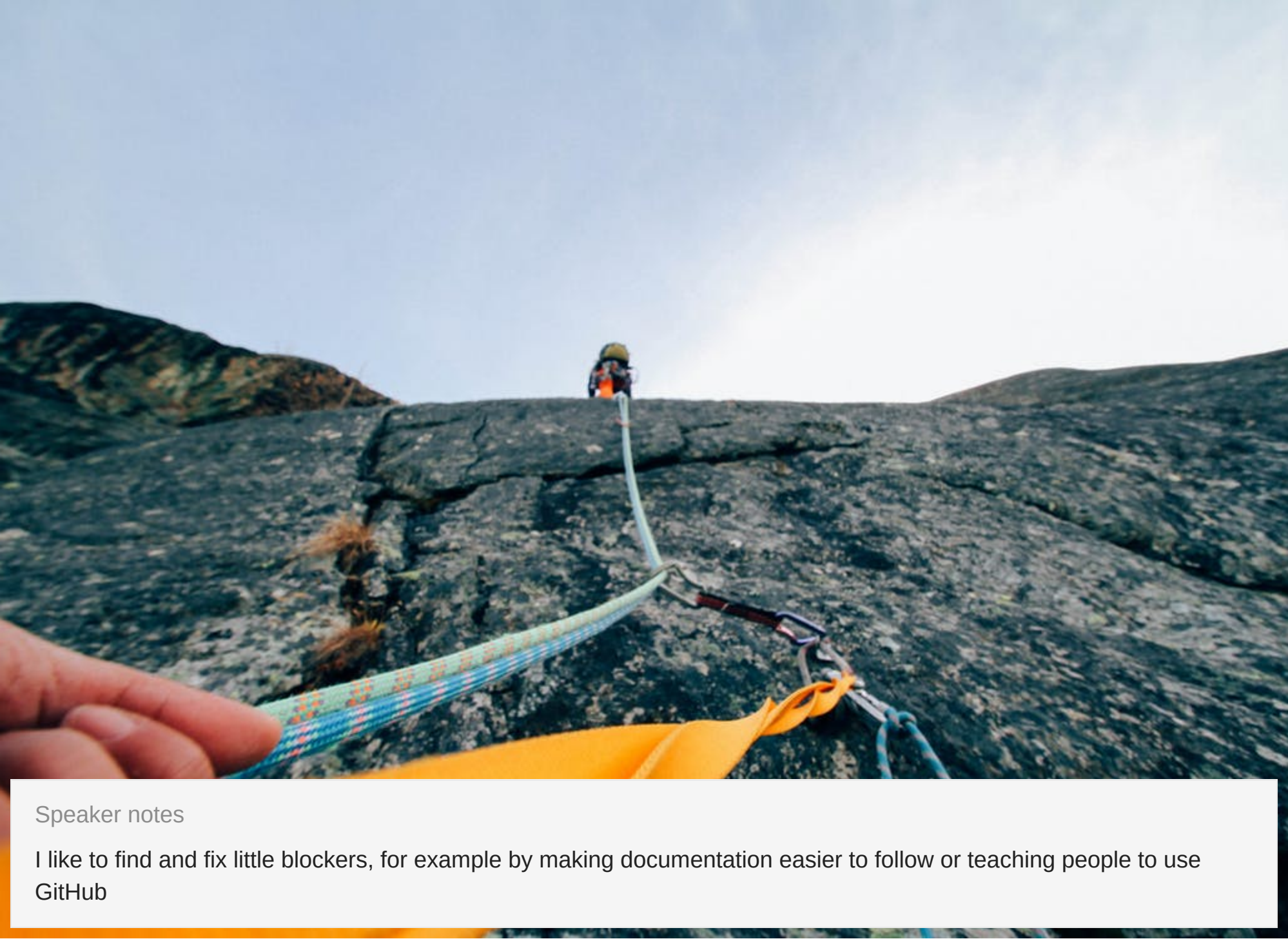
Jo Cook, Astun Technology

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

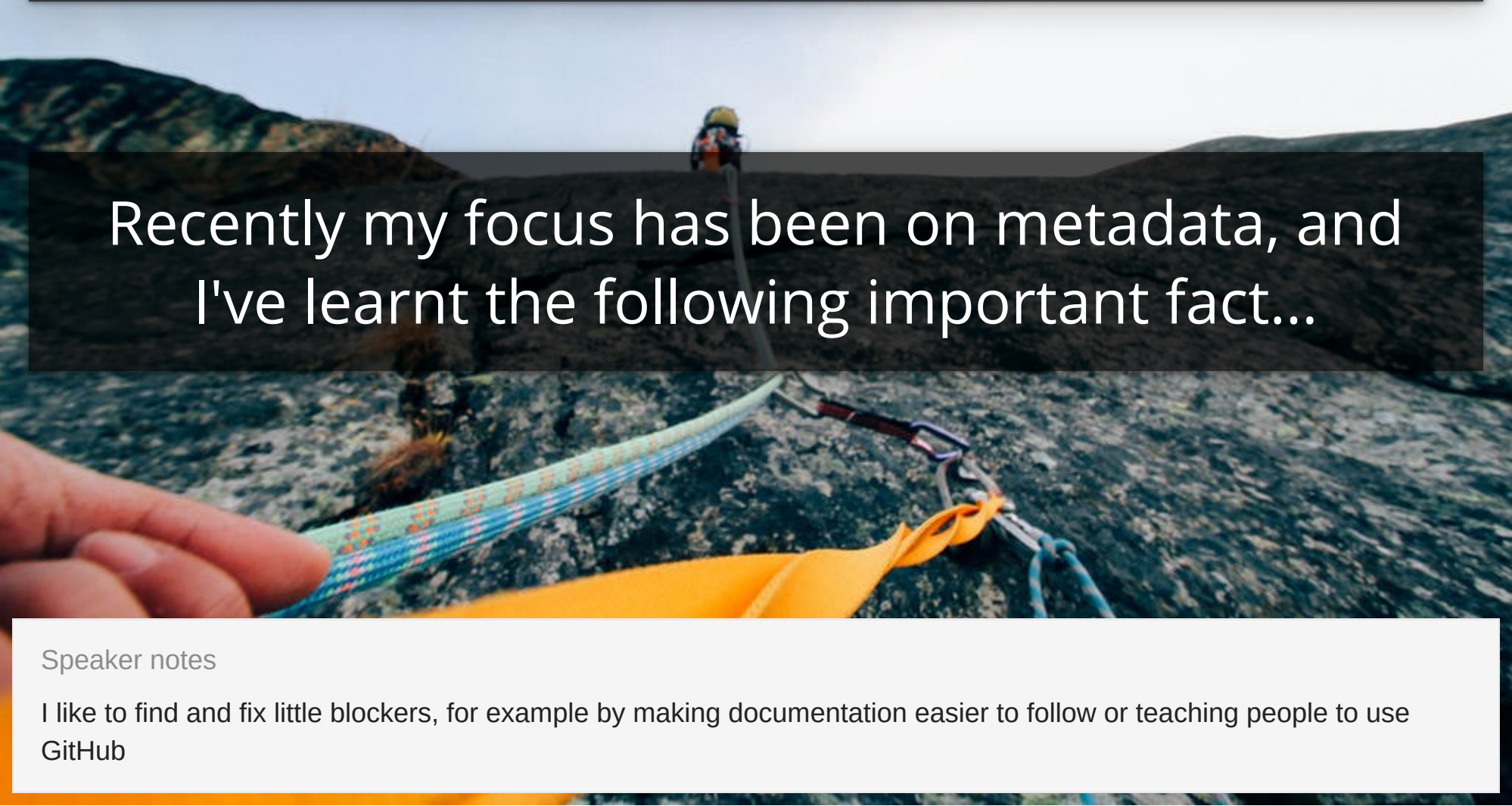
I like to find and fix little blockers, for example by making documentation easier to follow or teaching people to use GitHub

I'm not a coder! My passion is enabling others to do their job as easily as possible, preferrably 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, preferrably 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

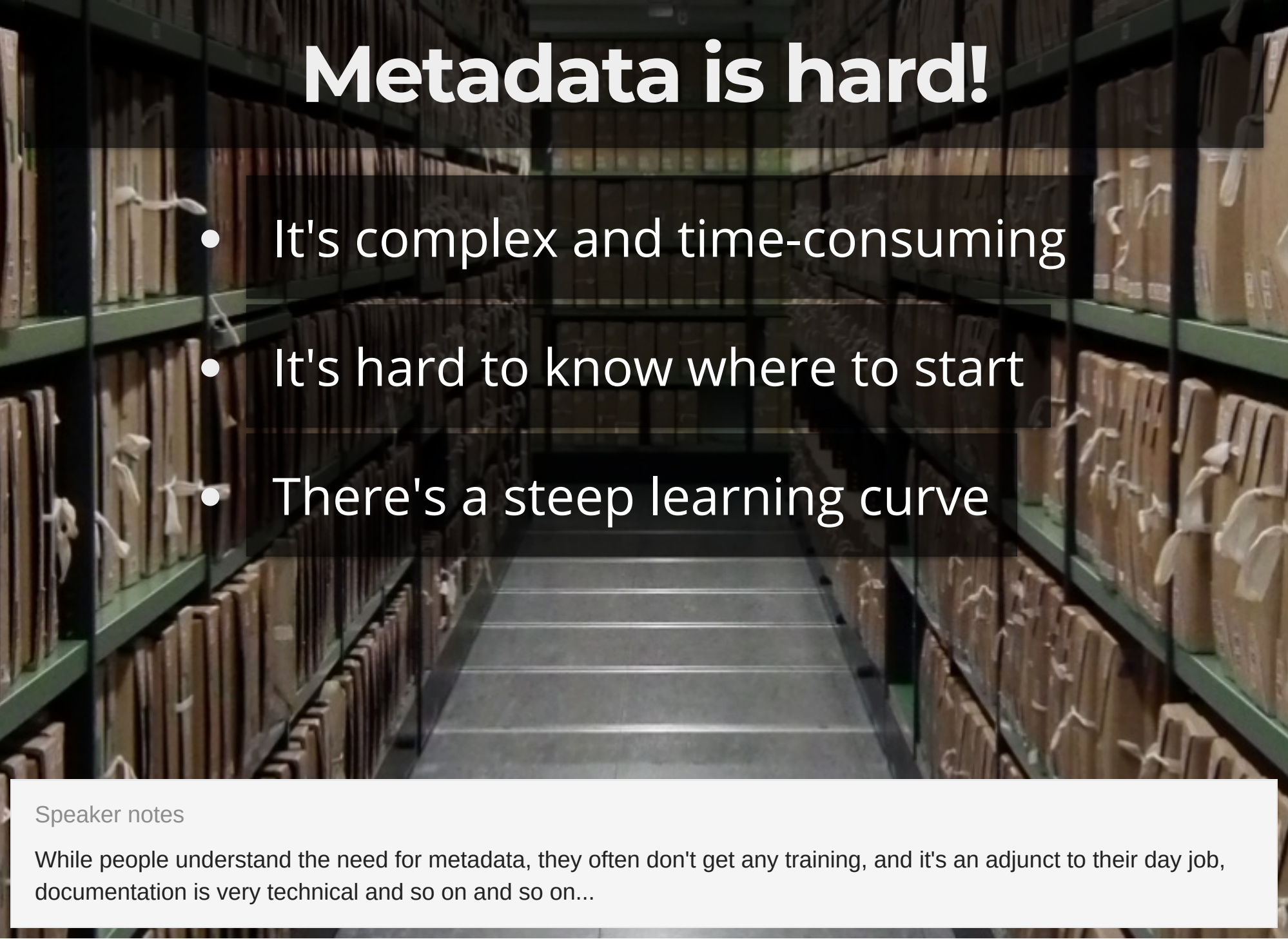
While people understand the need for metadata, they often don't get any training, and it's an adjunct to their day job, documentation is very technical and so on and so on...

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!

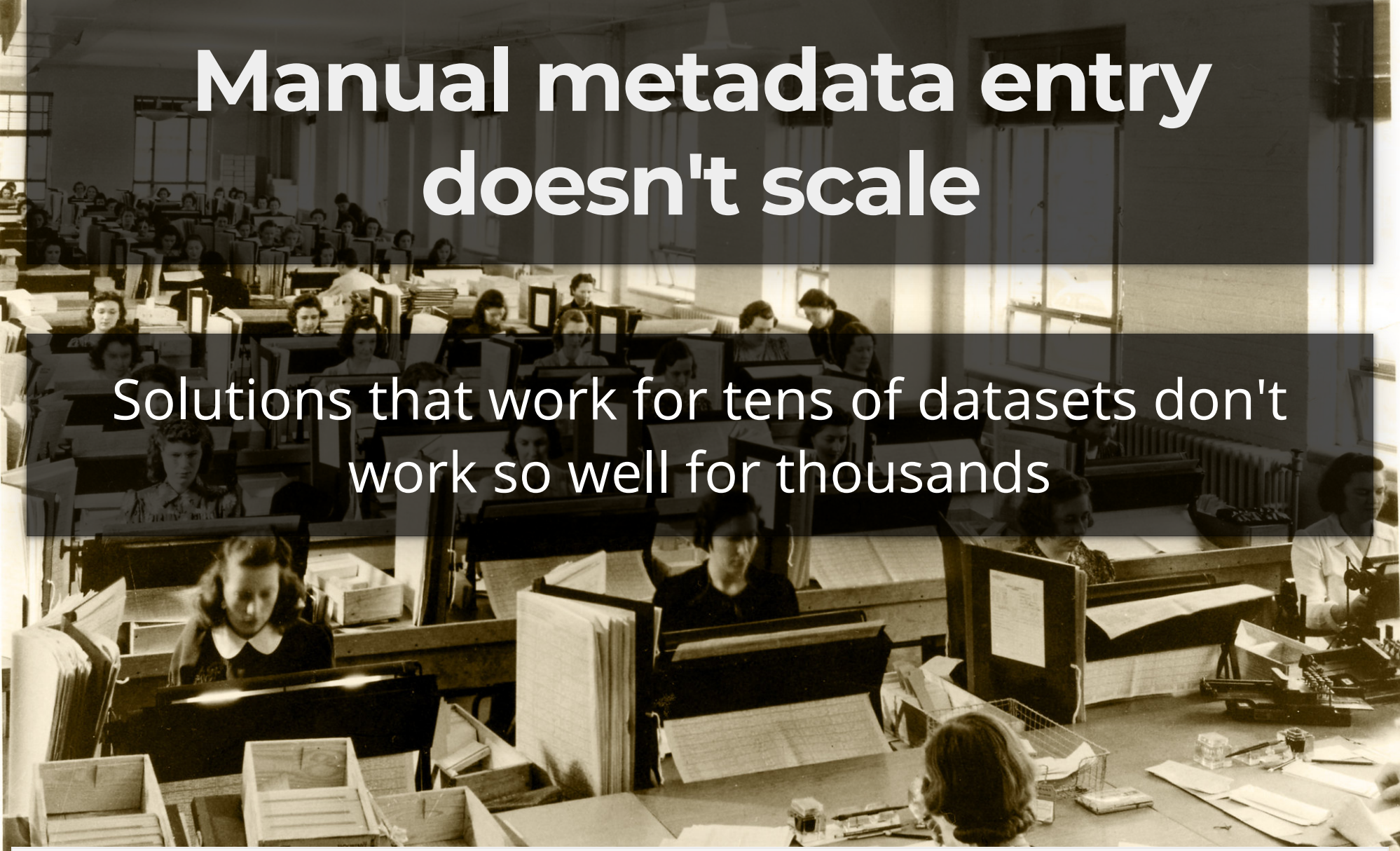
- It's complex and time-consuming
- It's hard to know where to start
- There's a steep learning curve

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

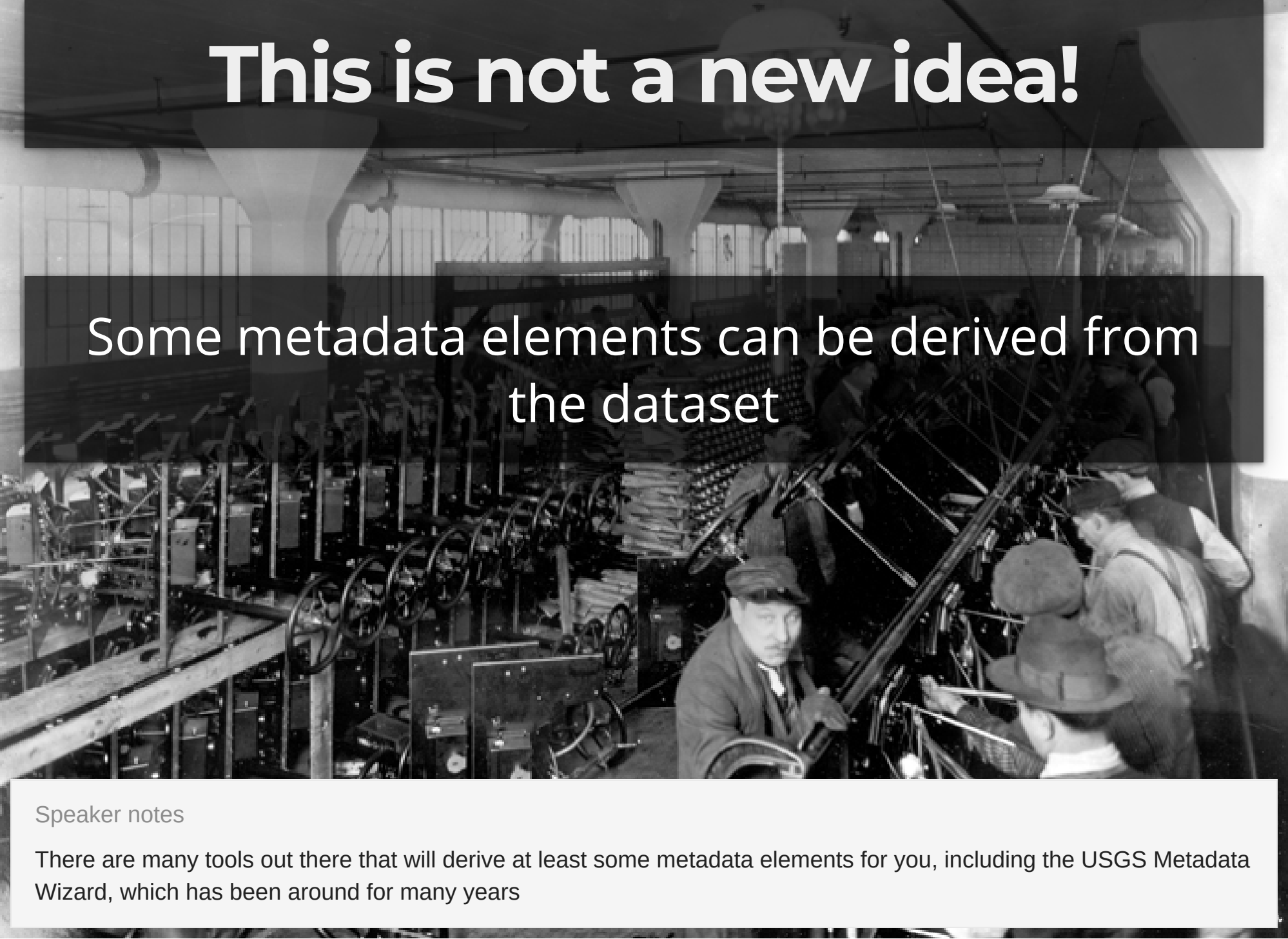


Speaker notes

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

Automate all the things!

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 Software for 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

Using Free and Open Source GIS to Automatically Create Standard Based Spatial Metadata in Academia - First Investigations

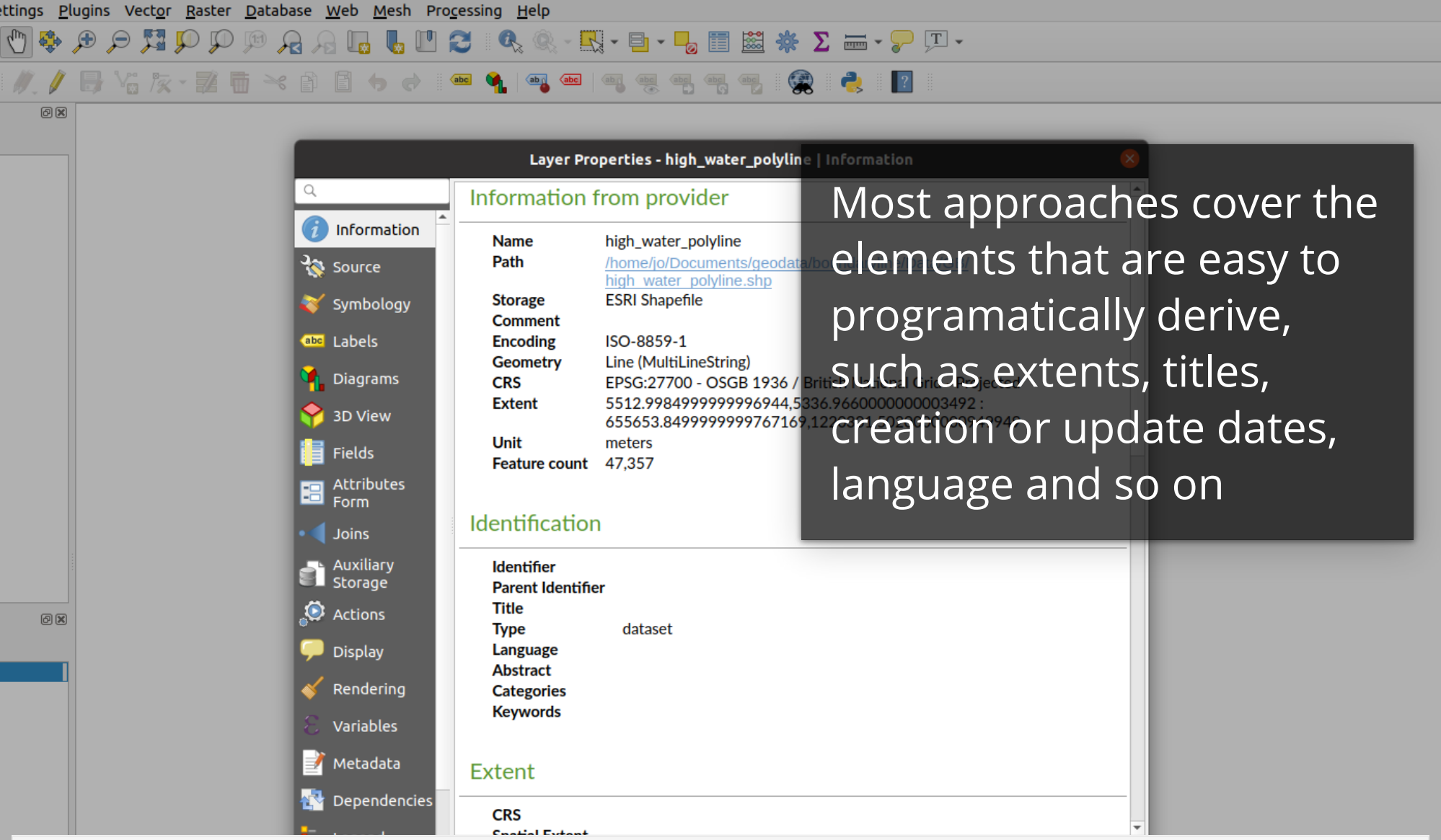
Claire Ellul
University College London, United Kingdom

Nart Tamash
University College London, United Kingdom

Feng Xian
University College London, United Kingdom

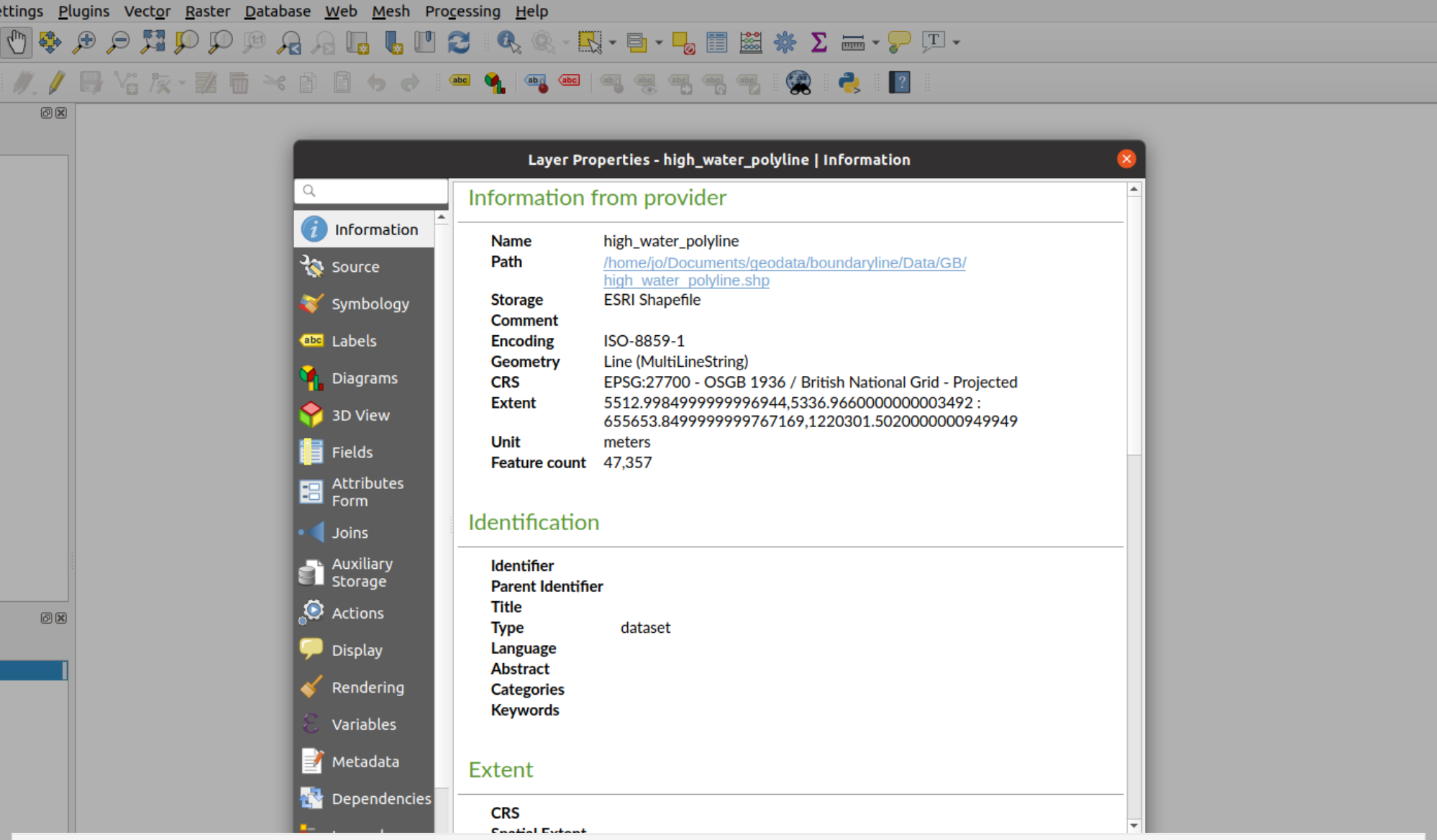
Speaker notes

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



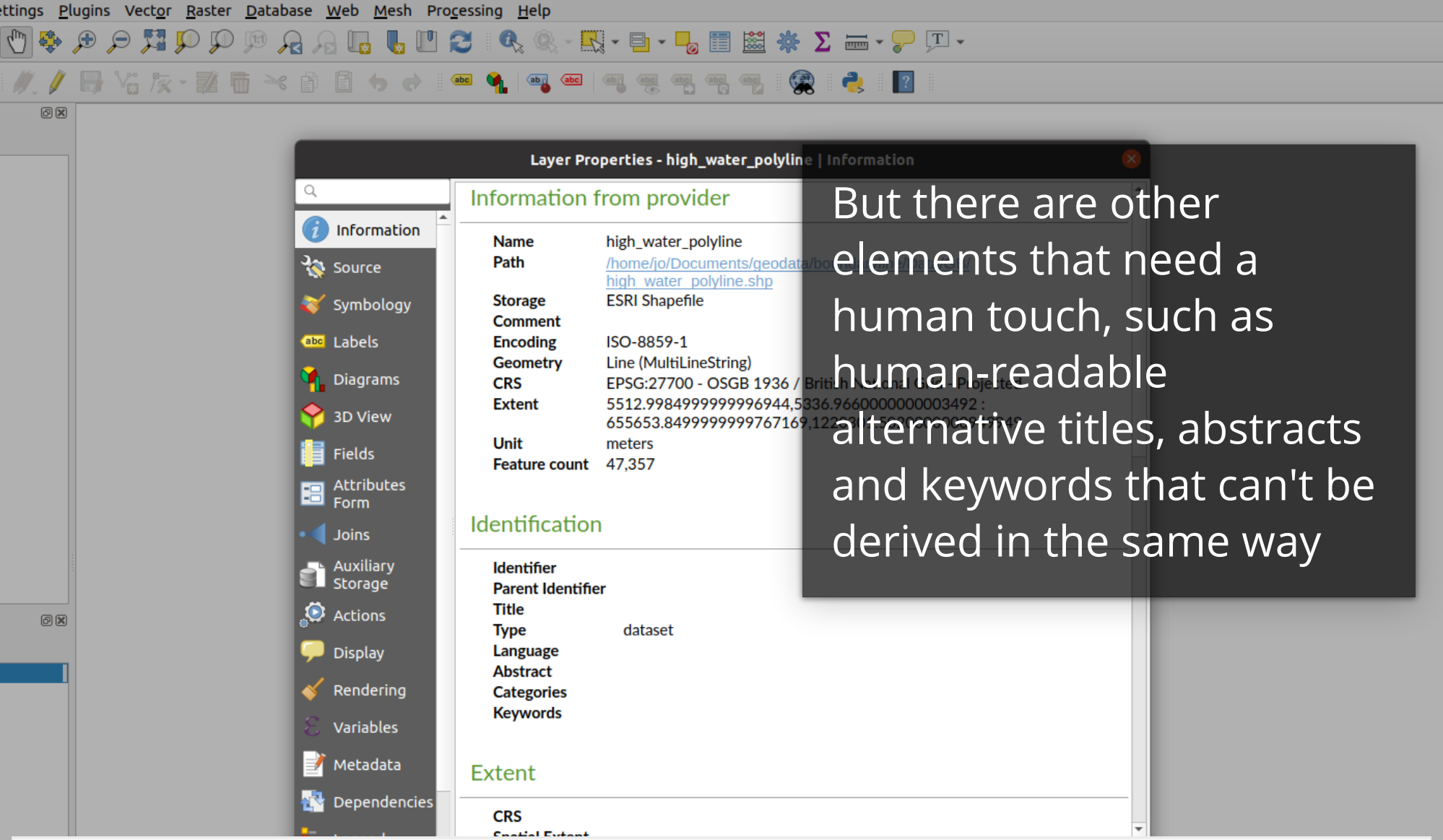
Speaker notes

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)



Speaker notes

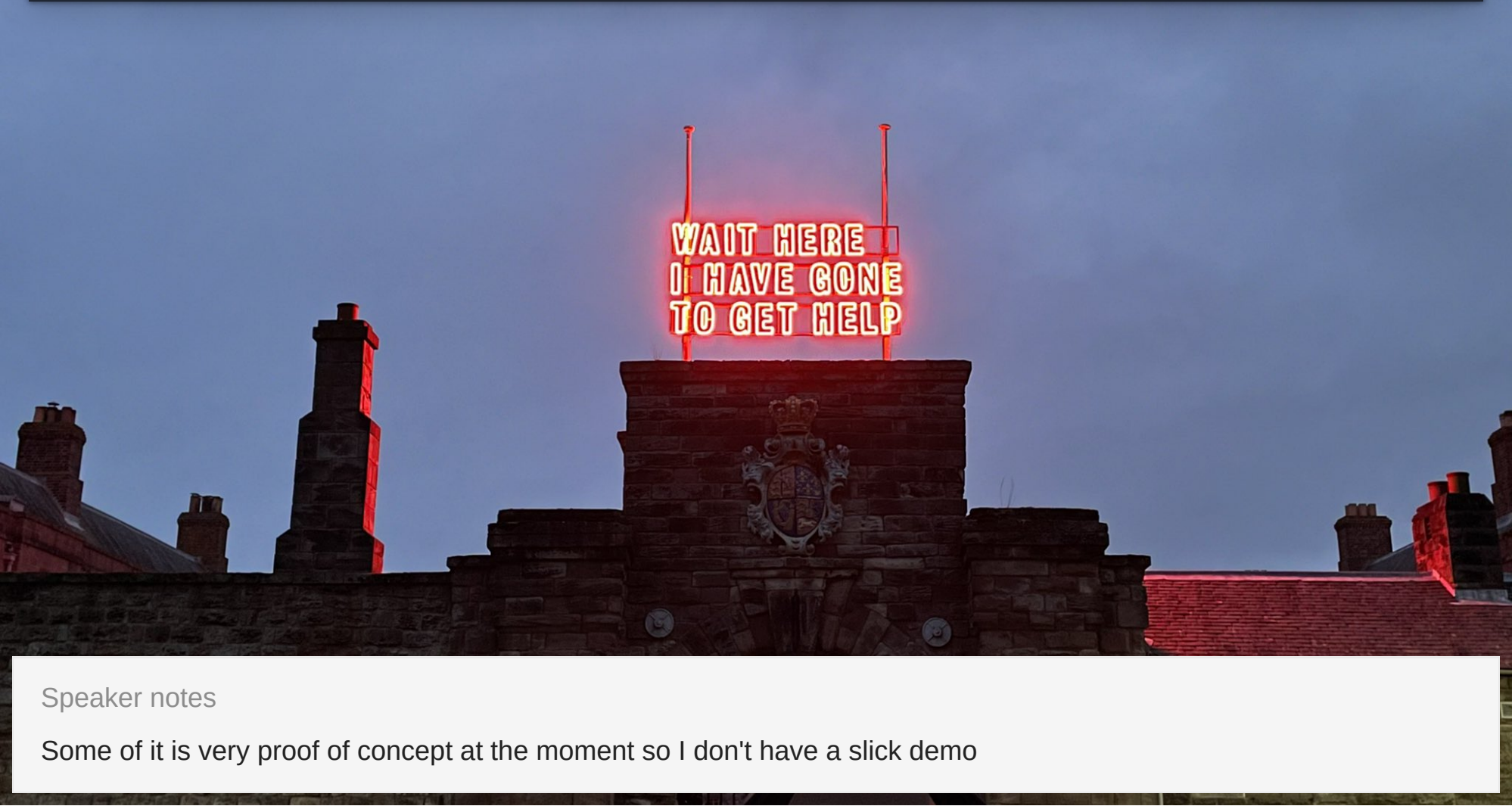
From personal observation, it's these manual elements that cause the most trouble and cause metadata projects to flounder or sometimes fail all together. These are increasingly important for data discovery and sharing (eg linked data)



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"



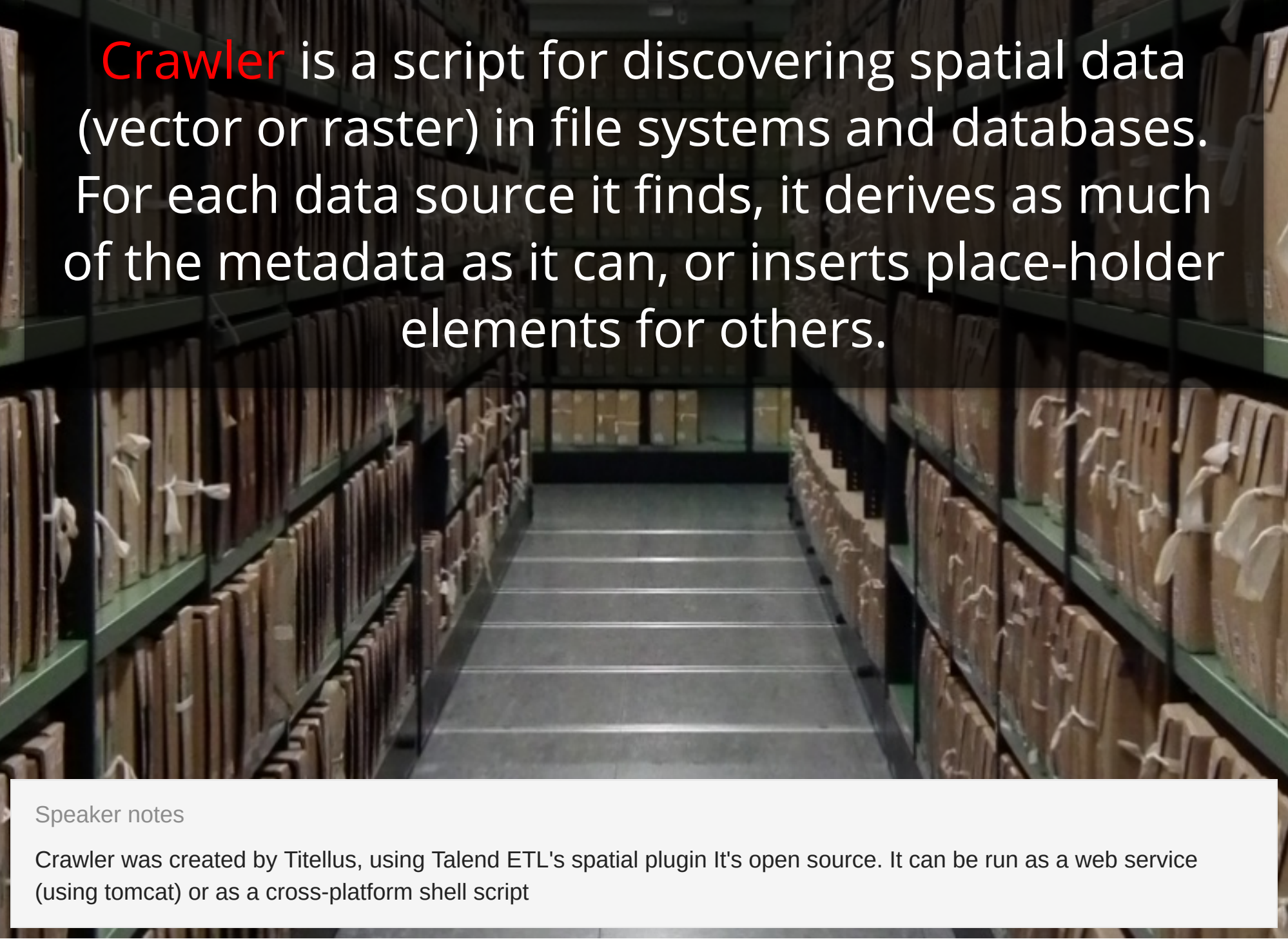
Speaker notes

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



Speaker notes

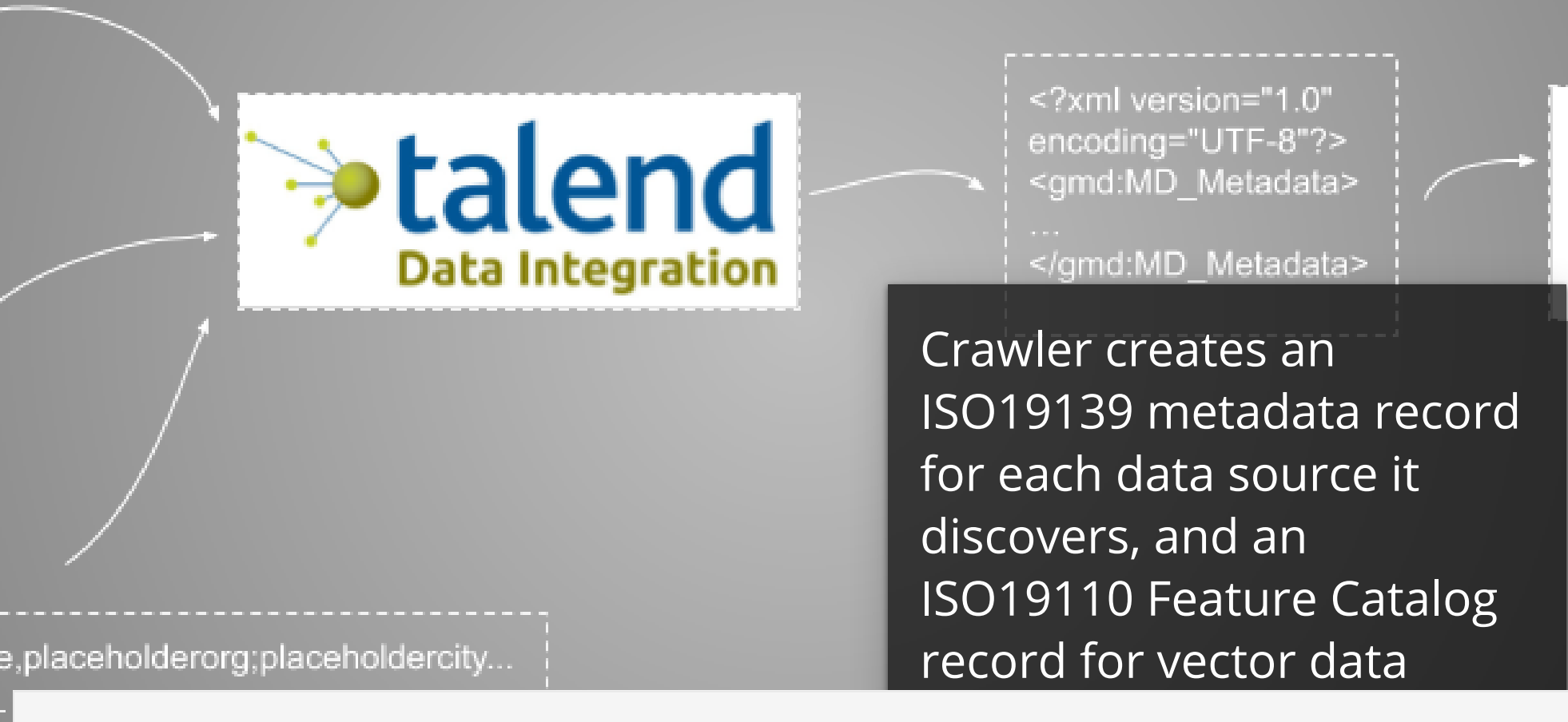
Crawler was created by Titellus, using Talend ETL's spatial plugin It's open source. It can be run as a web service (using tomcat) or as a cross-platform shell script

Step 1: From data to placeholder Metadata

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

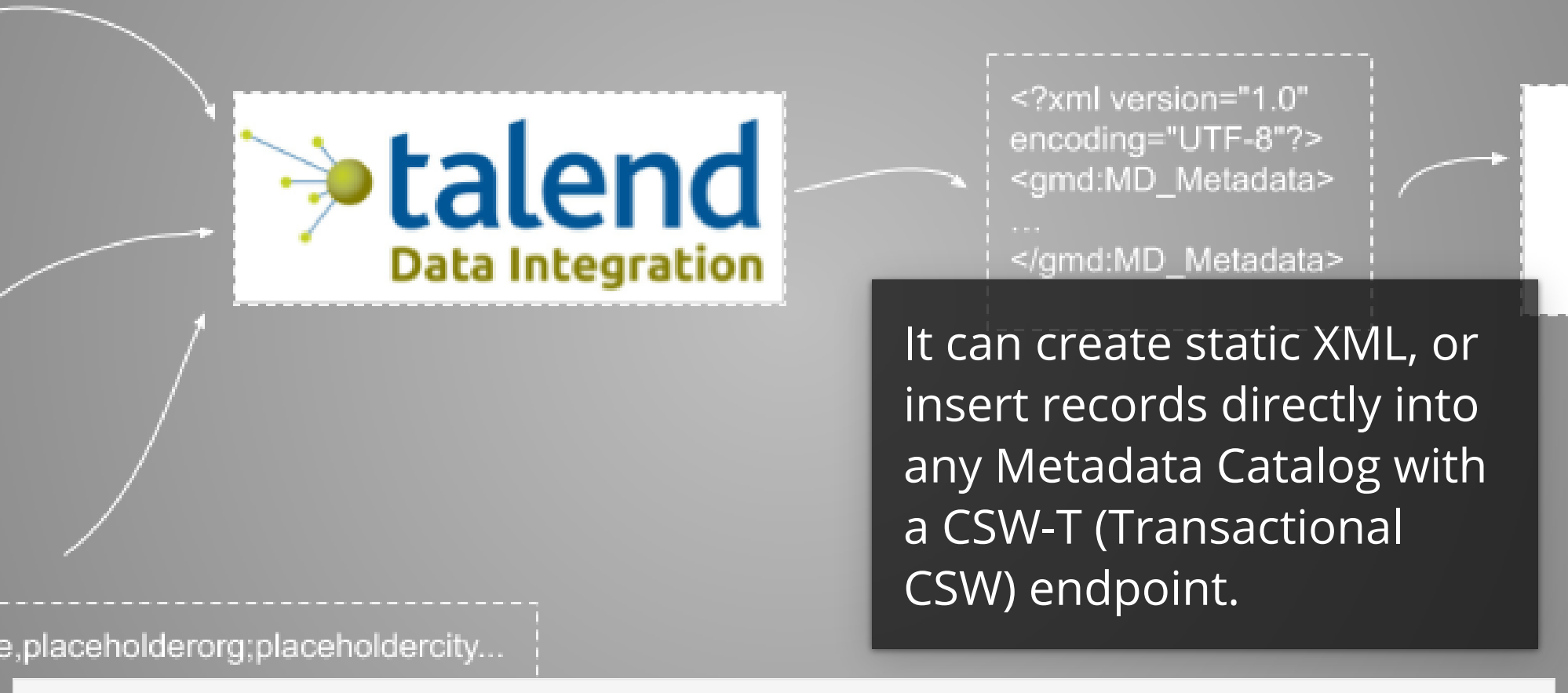


Crawler creates an ISO19139 metadata record for each data source it discovers, and an ISO19110 Feature Catalog record for vector data

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



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/

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)	

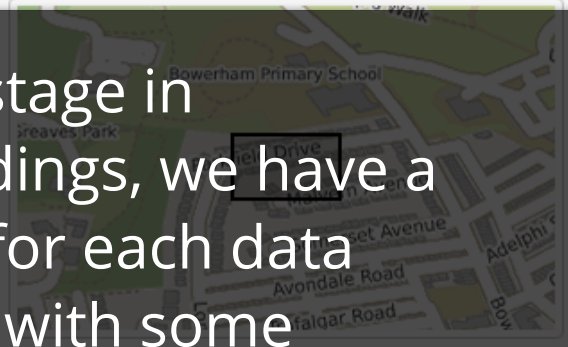
Feature catalog

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

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

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

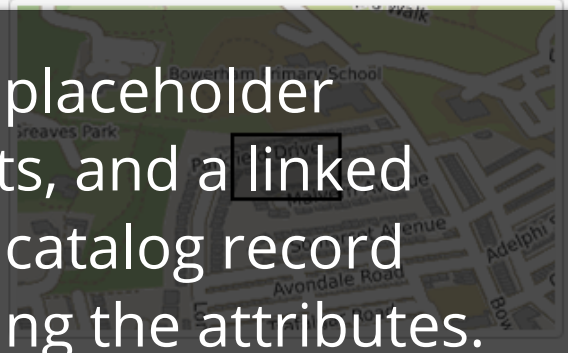
...some placeholder elements, and a linked feature catalog record describing the attributes.

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/

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)	

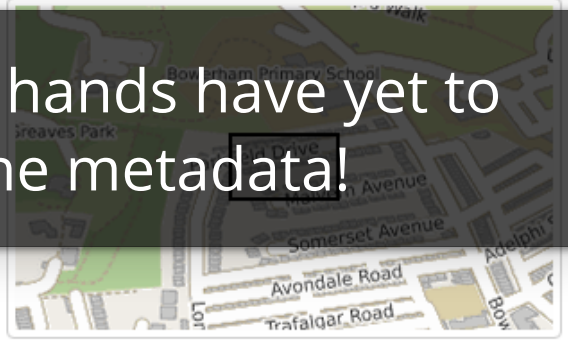
Feature catalog

Overview

No ratings ★

See all feedback | Add your review

Spatial extent



Human hands have yet to touch the metadata!

Temporal extent

Creation date
2000-01-01

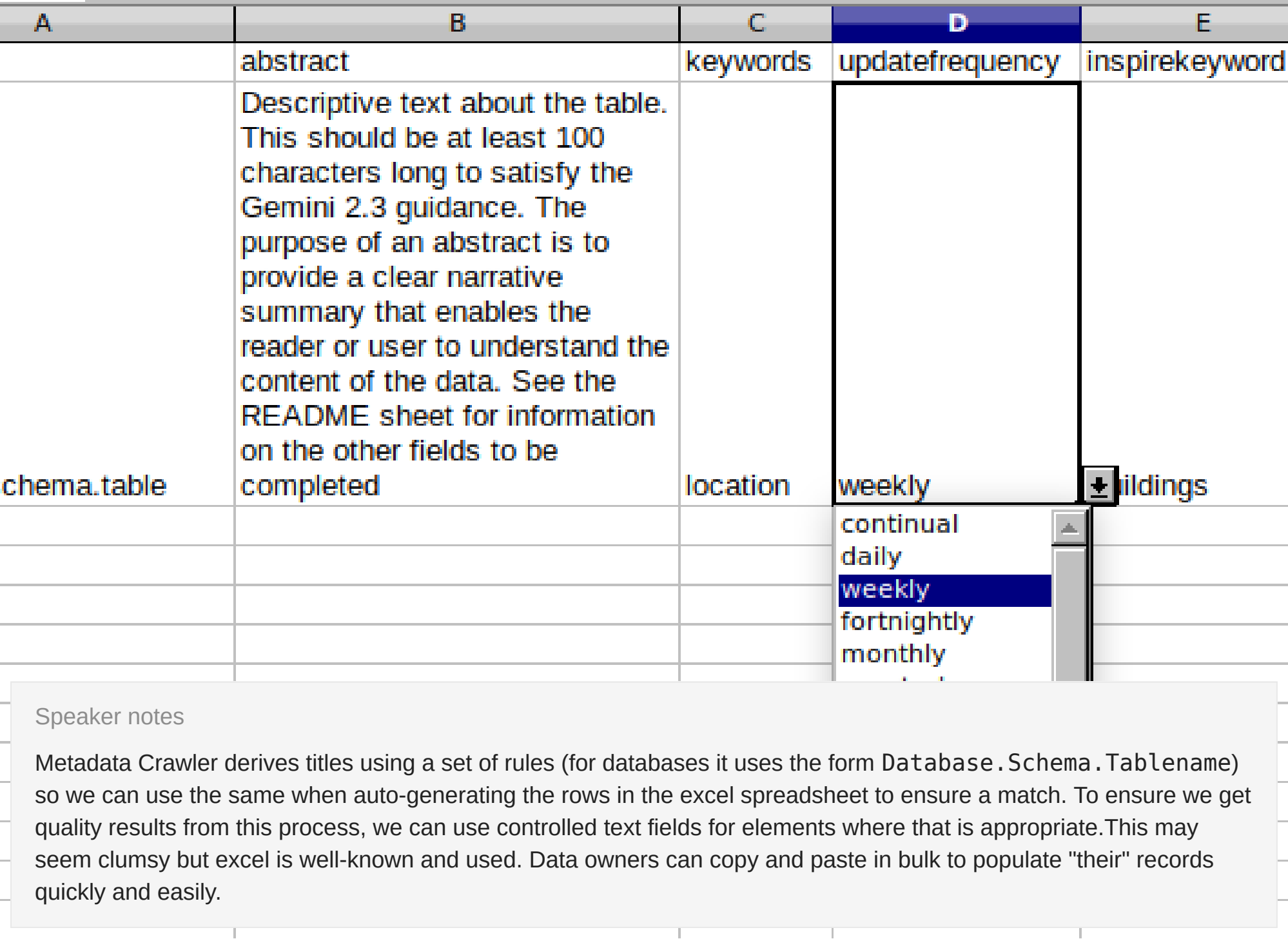
Revision date
2020-06-15

Period
Sat Jan 01 2000 00:00:00 GMT+0000 ⏩ Sat Jan 01 2000 00:00:00 GMT+0000

Provided by

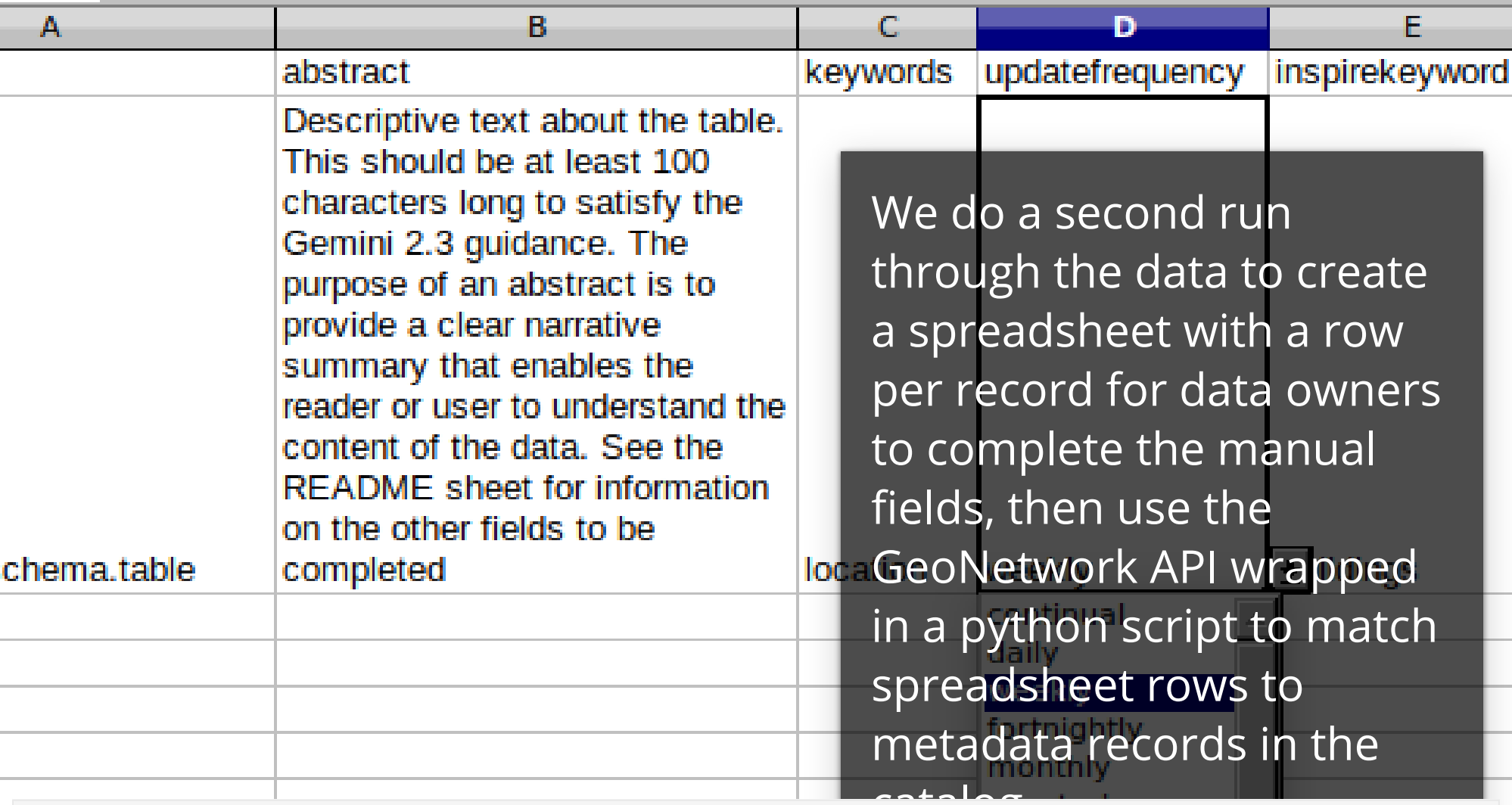
Part Two: Ye Olde Approache

medievalfragments.wordpress.com



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.



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

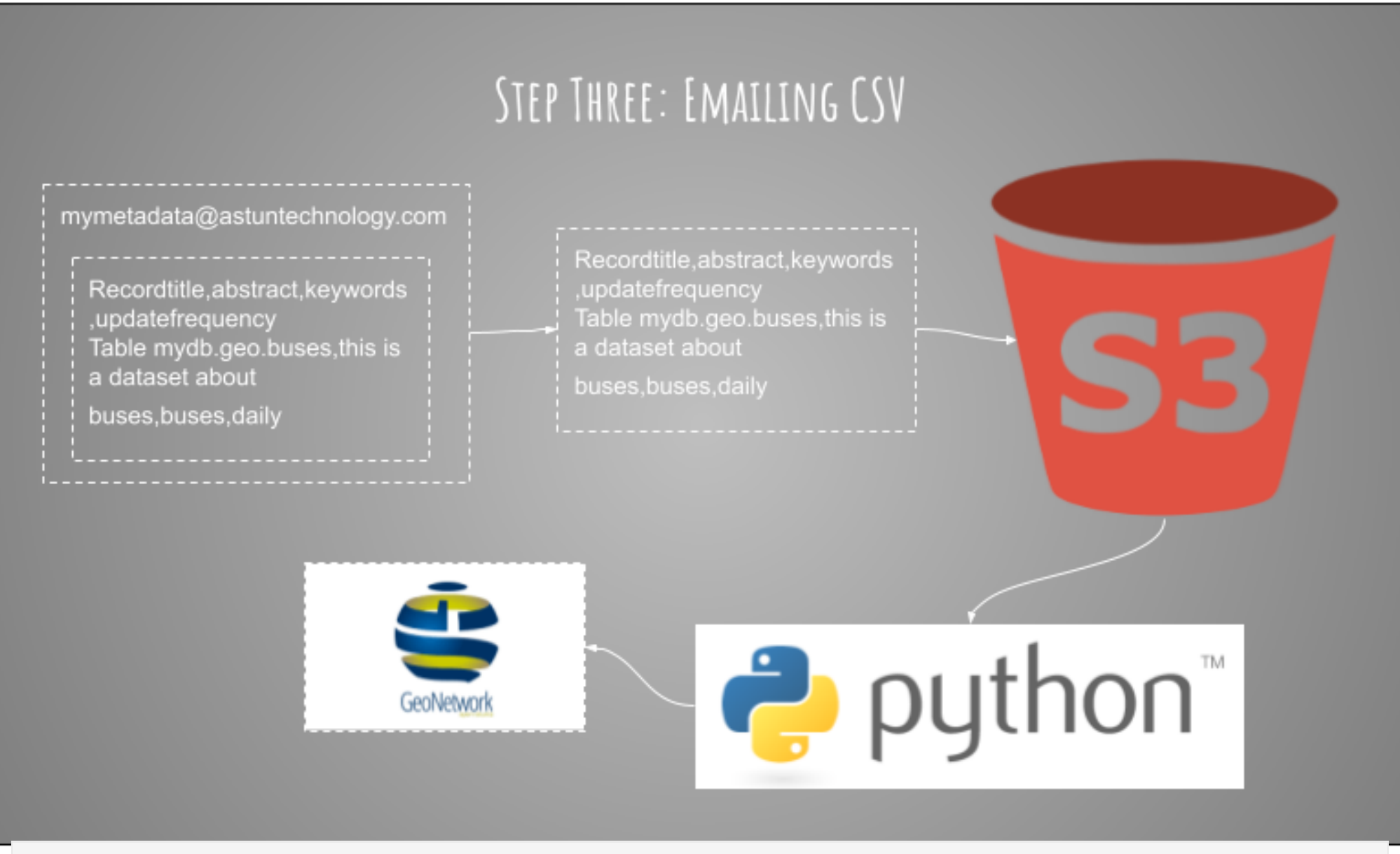
```
le,abstract,keywords,updatefrequency
db.geo.buses,this is a dataset about buses,buses,daily
```

```
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

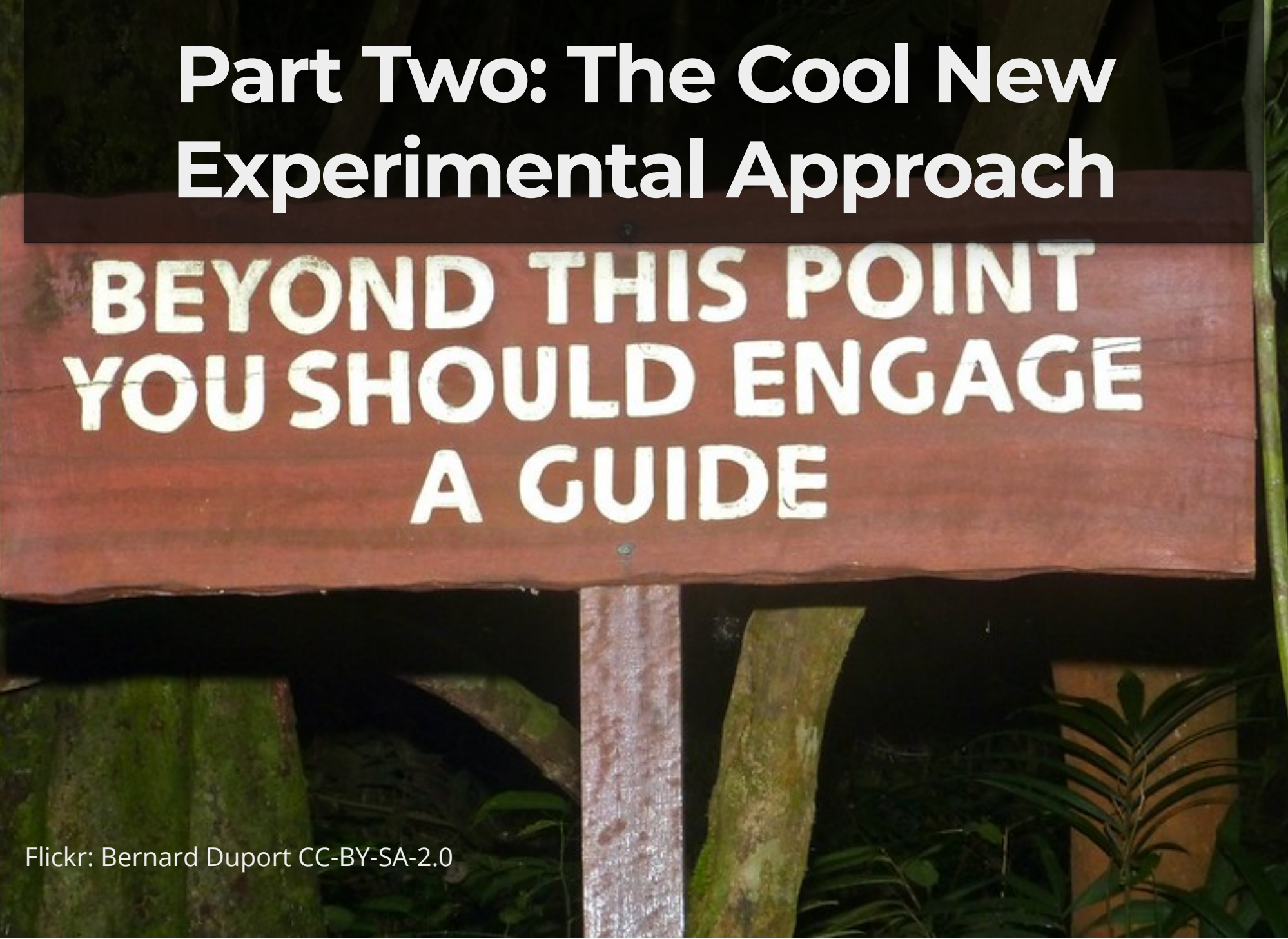
Even though GeoNetwork is hosted on AWS in the cloud, we can let users run these scripts using environments such as Cloud9

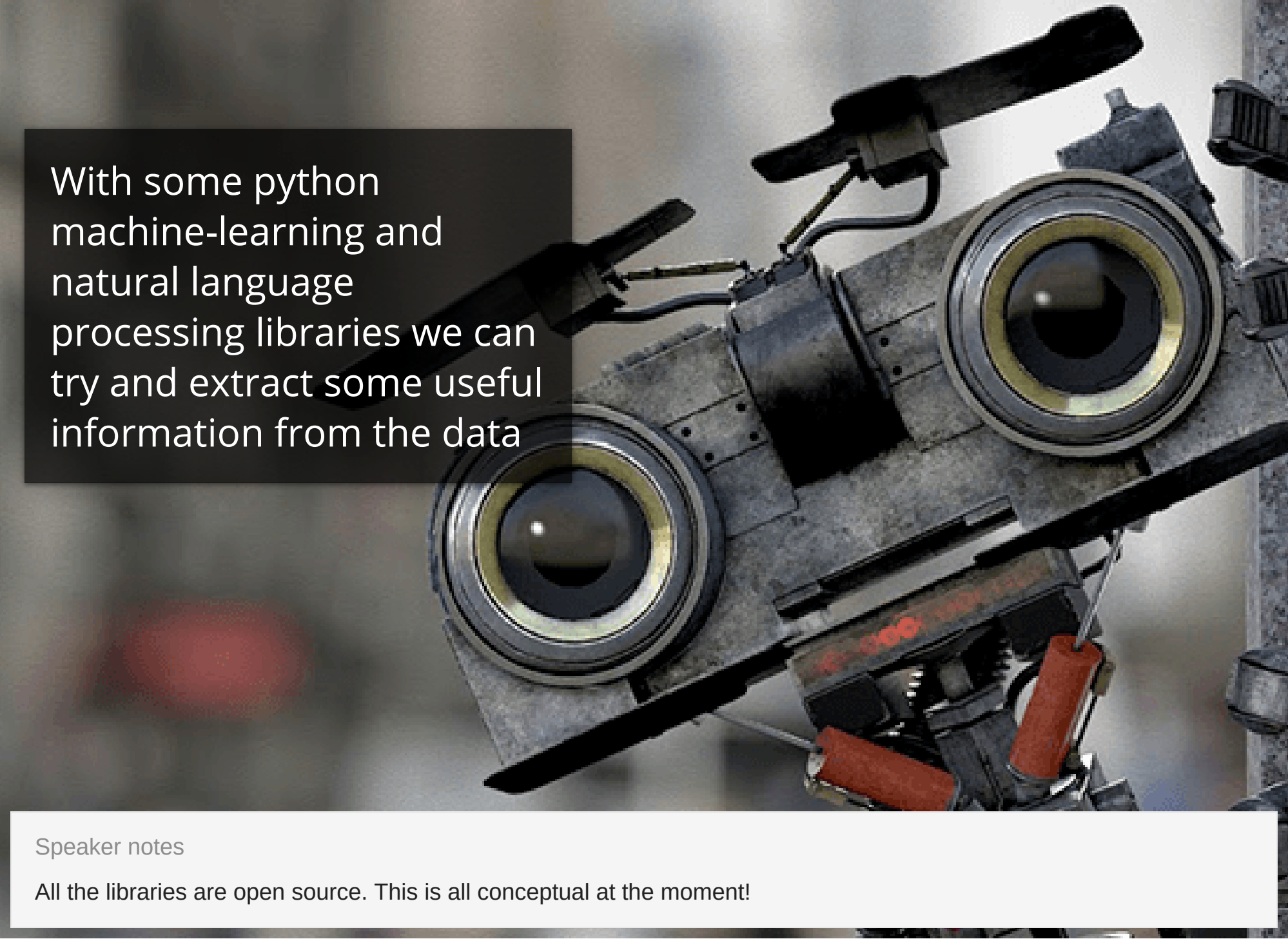


Speaker notes

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





Speaker notes

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

Machine-Learning Step One

Spatial Dataset

Geometries

Extent

GeoPanda

Reverse geocode centroid

Extract geographic keywords

Text Attributes (CSV)

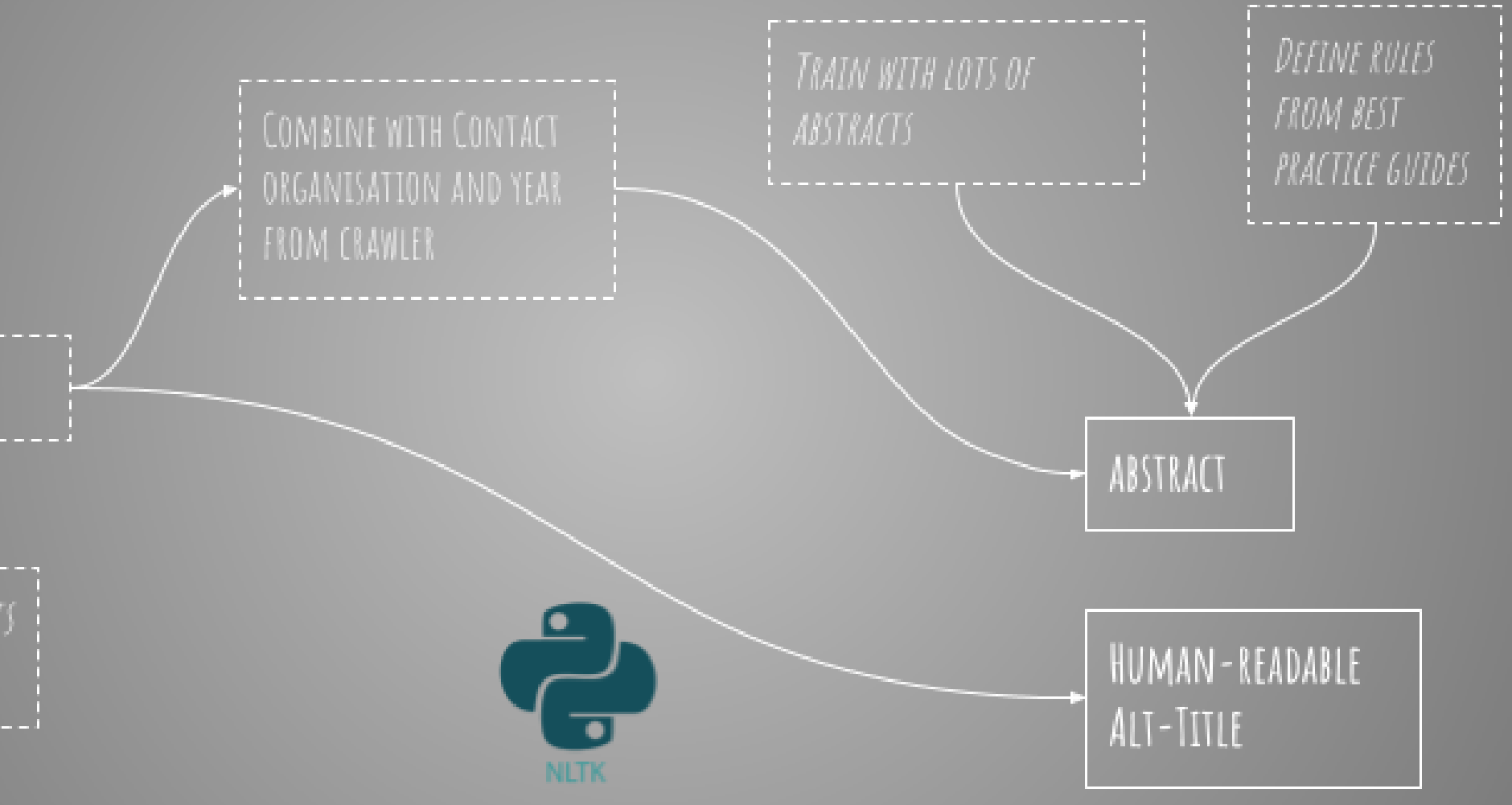
Extract keywords

KeyBERT

rank

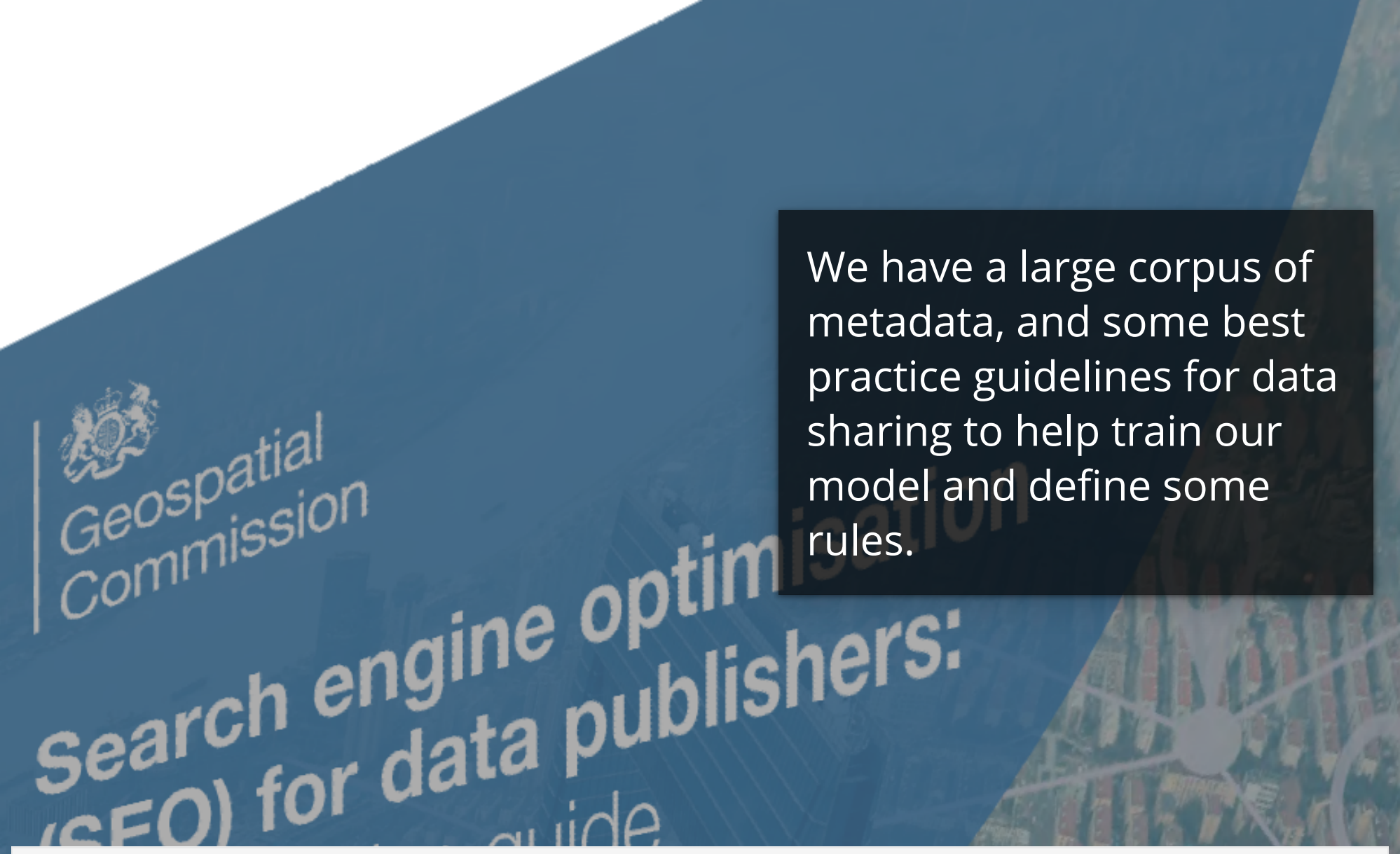
Cf with codelists

Machine-Learning Step two



Speaker notes

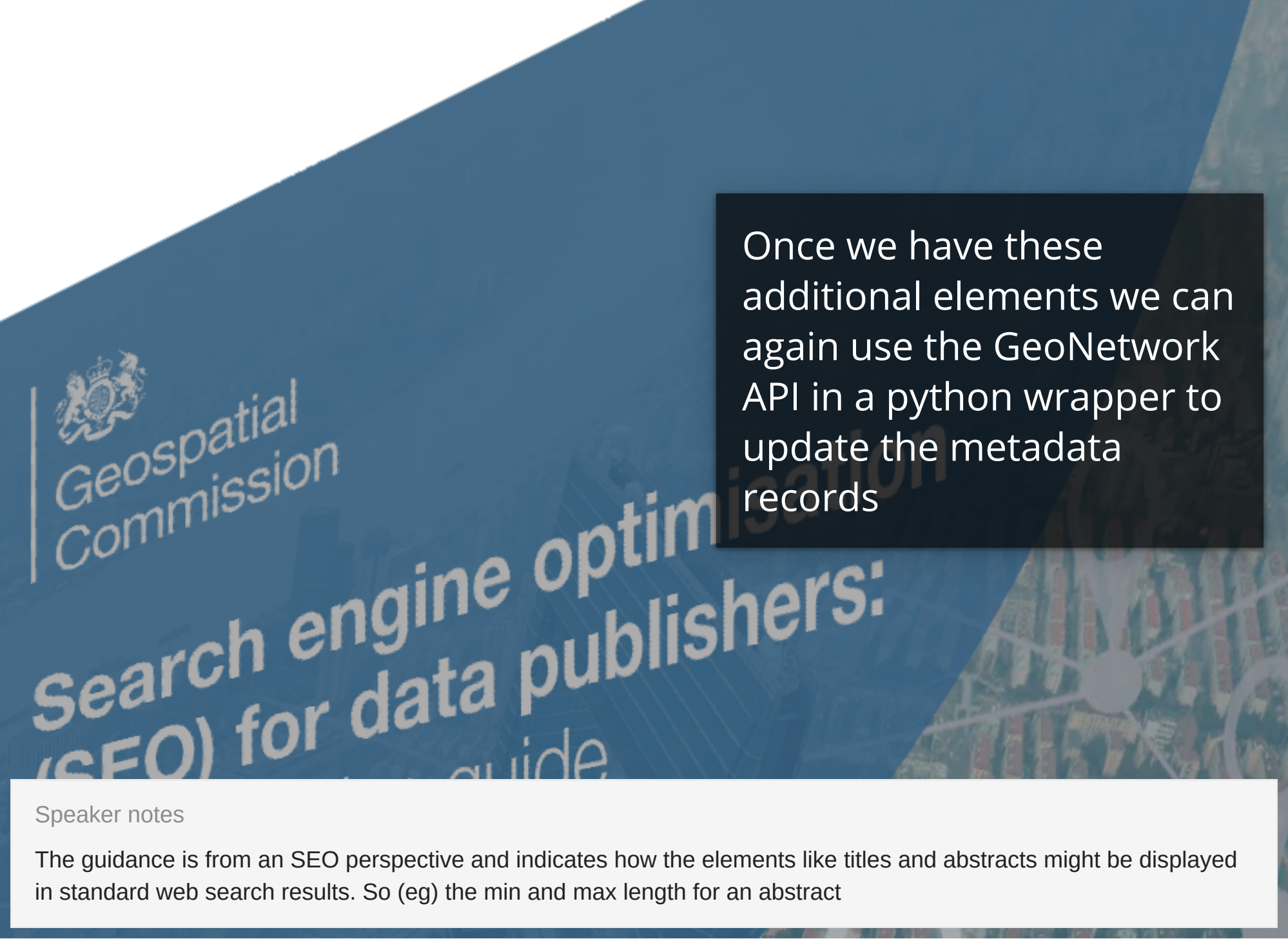
The guidance is from an SEO perspective and indicates how the elements like titles and abstracts might be displayed in standard web search results. So (eg) the min and max length for an abstract



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

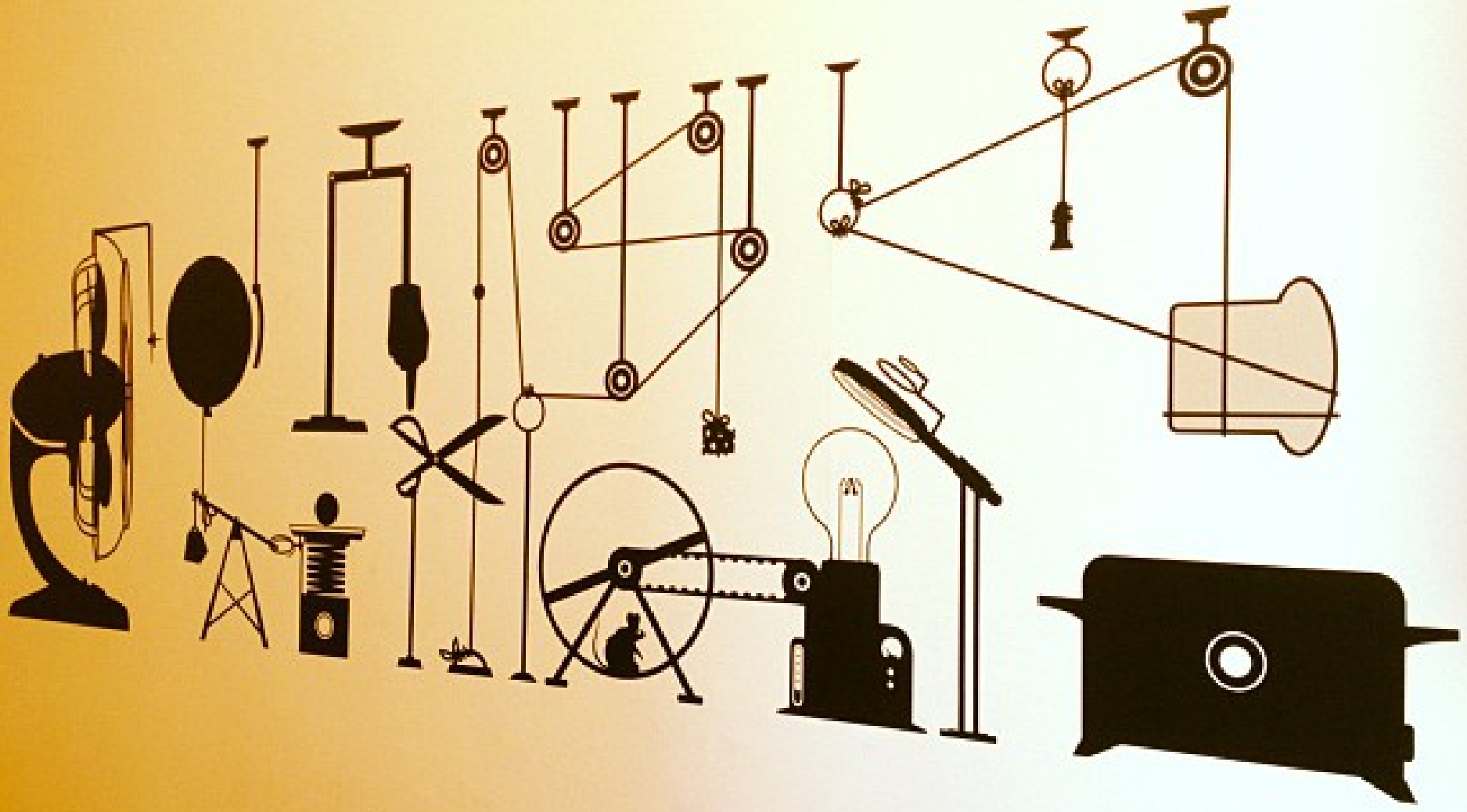
The guidance is from an SEO perspective and indicates how the elements like titles and abstracts might be displayed in standard web search results. So (eg) the min and max length for an abstract



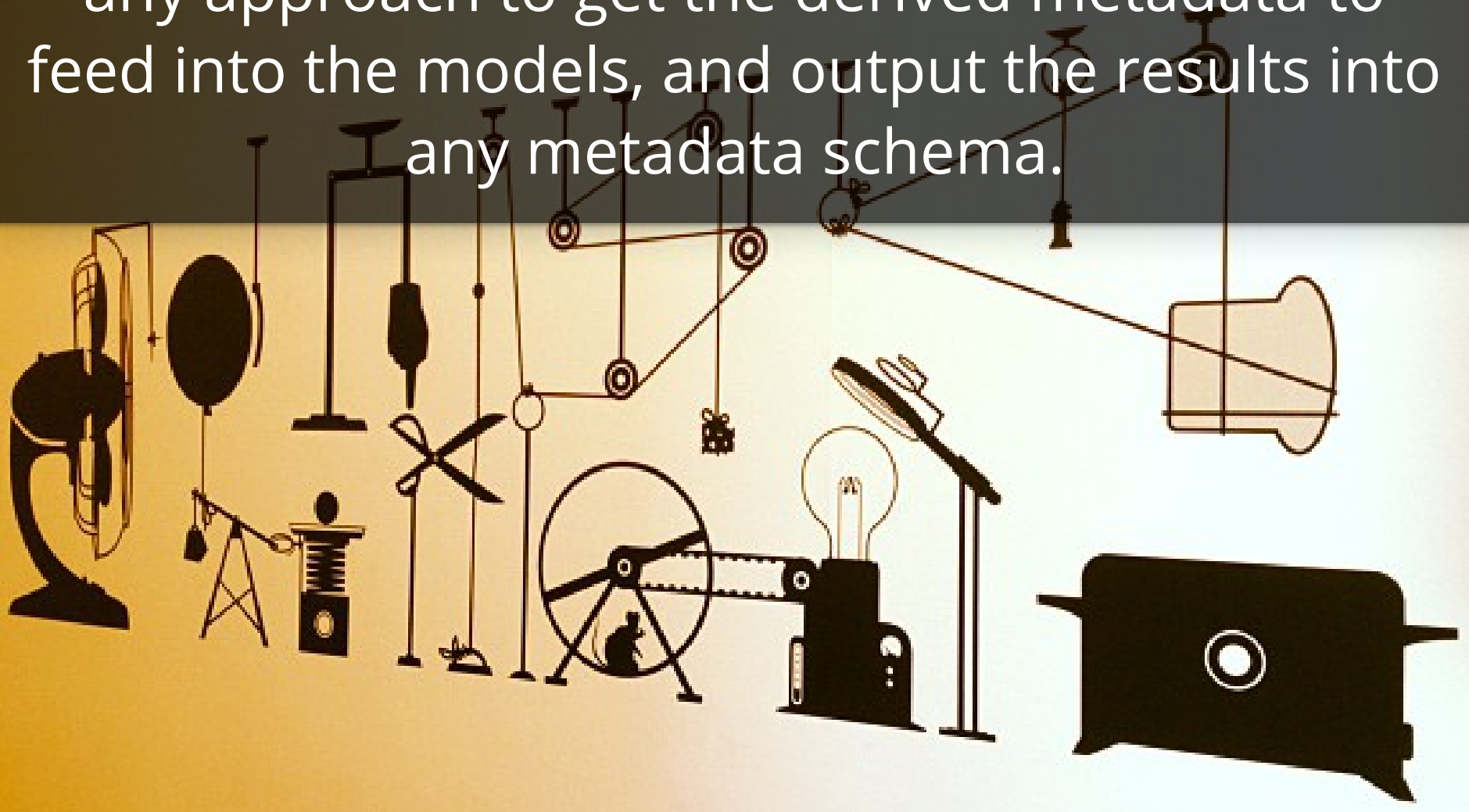
Once we have these additional elements we can again use the GeoNetwork API in a python wrapper to update the metadata records

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



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!

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!

Flickr: Kevan CC BY 2.0

Useful References

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

Thank You!

Jo Cook, Astun Technology

jocook@astuntechnology.com